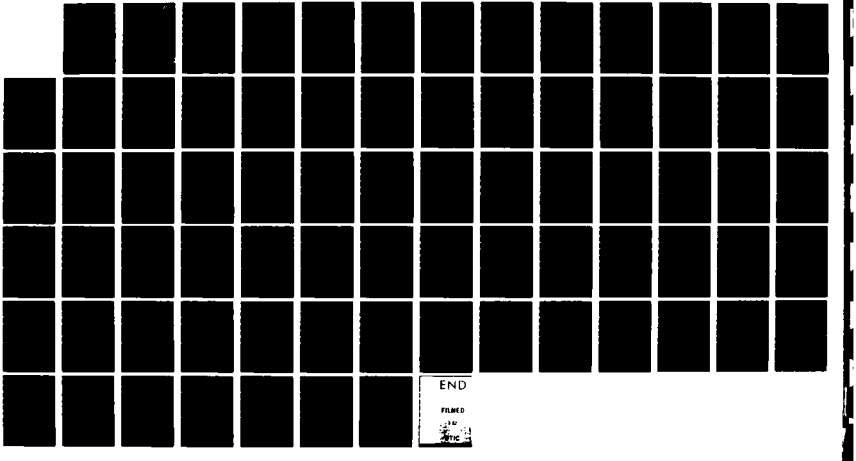
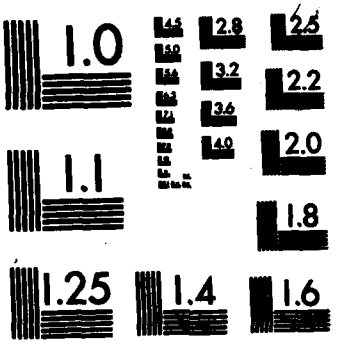
AD-A126 762 MASS SPECTROMETRIC RAPID DIAGNOSIS OF INFECTIOUS DISEASES(U) STATE UNIV OF NEW YORK AT BUFFALO SCHOOL OF MEDICINE M ANBAR ET AL. JAN 80 DAMD17-78-C-8035 1/1

UNCLASSIFIED F/G 6/5 NL





MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

ANNUAL PROGRESS REPORT

Michael Anbar
James McReynolds
Michael Scanlon
Robert Abbott
William Rieth
Lev Verkh

January 1980

MASS SPECTROMETRIC RAPID DIAGNOSIS OF INFECTIOUS DISEASES

Supported by

U.S. ARMY MEDICAL RESEARCH AND DEVELOPMENT COMMAND
Fort Detrick, Maryland

Contract DAMD17-78-C-8035

School of Medicine
State University of New York at Buffalo
Buffalo, New York

UNCLASSIFIED

Department of Biophysical Sciences
School of Medicine
State University of New York at Buffalo
Buffalo, New York 14214

a Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)	2a. REPORT SECURITY CLASSIFICATION
	2b. GROUP

3. REPORT TITLE

MASS SPECTROMETRIC RAPID DIAGNOSIS OF INFECTIOUS DISEASES

4. DESCRIPTIVE NOTES (Type of report and inclusive dates) PROGRESS REPORT
second annual report for period of February 1, 1979 - January 31, 1980

5. AUTHOR(S) (First name, middle initial, last name)

Anbar, Michael
McReynolds, James
Scanlon, Michael
Abbott, Robert
Rieth, William
Verkh, Lev

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
January 28, 1980	73	

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
DAMD17-78-C-8035	
b. PROJECT NO.	
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	

10. DISTRIBUTION STATEMENT

Approved for public release, distribution unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	U.S. Army Medical Research & Development Command Ft. Detrick, Maryland

13. ABSTRACT

Using new chemical procedures, mass spectrometric instrumentation and appropriate computerized data analysis, the diagnosis of a number of infectious diseases, through the molecular weight profile of neutral metabolites urine, was demonstrated. Longitudinal studies on human volunteers infected with sandfly fever showed the appearance of a characteristic pattern prior to the onset of clinical symptoms which persisted some time after the symptoms have subsided. Tissue cultures infected with polio virus exhibited a characteristic pattern demonstrat able within 6 hrs from infection.

DD FORM 1 NOV 1473

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Clinical diagnosis, infectious diseases, metabolite profiles, urine, field ionization mass spectrometry, multi-component analysis, molecular weight profiles.						

1. INTRODUCTION

This is the second annual progress report on a research program entitled, "MASS SPECTROMETRIC RAPID DIAGNOSIS OF INFECTIOUS DISEASES", under Contract No. DAMD 177808035, sponsored by the Department of Army, U.S. Army Medical Research and Development Command, Fort Detrick, Maryland. This progress report covers the period from February 1, 1979 through January 31, 1980.

The capability of making a rapid and reliable diagnosis of infectious diseases at any early stage and at low cost would be of especially great value to the military where large numbers of army personnel are stationed in confined areas and their continuing health is crucial to carrying out their objectives. Early and reliable diagnosis of an infectious disease could prevent the spread of disease to large groups of military and civilian personnel.

Multicomponent analysis may be used to identify in the host's urine characteristic metabolic patterns associated with infection in general, with bacterial or viral infection, or with specific infections. The multiscan mass spectrometric method offers four types of uses in the diagnosis of infectious diseases. First, multicomponent analysis by mass spectrometry may be used as a diagnostic tool, allowing diagnosis of an infection during the incubation period, thus facilitating timely isolation and appropriate treatment. Second, the same technique may be used for the identification of bacteria and viruses in vitro. Third, the characteristic components identified by the pattern recognition approach, can be chemically characterized by the FI-CID technique and lead to an understanding of the biochemical nature of the host's reaction. Fourth, a small number of metabolites identified by mass spectrometry could then be determined by non mass spectrometric analytical techniques (e.g., glc, hplc, or specific fluorometric ceterminants) and advantageously used for routine early diagnosis.

During the second year of this second phase, we have achieved a number of critical objectives. After further improvements and optimization of the sample preparation techniques and after finding optimal conditions for mass spectrometric analysis (utilizing a double focussing configuration), we have devoted a substantial effort to compare and critically analyze alternative statistical multivariable diagnostic procedures. However, most of the effort has been devoted this year to the analysis of clinical samples and thus to the evaluation of the potential usefulness of our methodology as a clinical diagnostic technique. First we have demonstrated a high degree of success in separating a number of different groups of patients by the metabolic profiles of their urines. These included patients with alcoholic liver disease, children with pneumonia and children with virus induced diarrhea; the patients could be readily differentiated by comparison with healthy subjects of the corresponding age group with practically no false-positives or false-negatives. Next we have carried out a longitudinal study on urine samples obtained from groups of volunteer subjects vaccinated with live virus of sandfly fever and dengue fever and followed up for a number of weeks. Although this study is not yet complete, it suggests that a diagnostic pattern associated with the infection can be detected before the onset of clinical symptoms and it subsides after their disappearance. Also differences in the rate of individual reaction to the infection could be shown.

In another study we have shown that the molecular weight profile of a human tissue culture medium exhibits significant changes when the cells are infected with a virus, within hours following infection. Human lung tissue cultures infected with polio mellitus demonstrated a diagnostically useful pattern 6 hours following exposure to the virus. This feasibility study is now being continued to establish the scope and limitations of this early detection technique.

In brief during the second year of this project preliminary experiments have shown that metabolic profiles of the host can be used to identify infected subjects, to differentiate between patients with different infections, to detect a viral infection prior to the onset of clinical symptoms and to demonstrate the infection some time after the clinical symptoms have subsided. Next we have shown that mass spectrometric multicomponent analysis can be a highly useful tool in the clinical laboratory by detecting the presence of a virus through its effect on the metabolism of tissue cultured cells. A diagnostic biochemical pattern seems to be distinguishable within a few hours which is significantly faster than by the presently used morphological changes. The coming third year of this project will be devoted to substantiating our findings on human host reactions and on early virus detection and identification in vitro.

To reiterate the above in a more explicit manner, during 1979 we have accomplished the following tasks:

1. Developed and tested sample sterilization and storage procedures.
2. Developed new, simpler sample preparation techniques including one to handle tissue culture media.
3. Improved on the mass spectrometric procedure.
4. Improved the interfacing with the INCOS-NOVA and the CYBER 173 computers.
5. Tested and evaluated different statistical analysis classification procedures.
6. Analysed numerous samples of children and adults with different pathological problems and demonstrated the efficiency of our diagnostic procedure.

7. Analysed samples of series of samples obtained from virus infected volunteers over a period of 3 to 4 weeks and demonstrated the appearance of a pathological pattern which disappeared after the disappearance of the clinical symptoms.
8. Analysed media of human tissue culture infected with polio virus 6 and 24 hours following exposure to the virus and demonstrated detectable changes at 6 hours.

2. EXPERIMENTAL TECHNIQUES AND PROCEDURES

The following is an updated description of our experimental techniques and procedures. Although some of the material described below was included in the previous report, significant changes have been introduced this year in a number of phases in the analytical procedures to warrant reporting anew the whole analytical procedure.

A. Sample Collection and Storage Procedure

Urine samples were collected by hospital staff in 12 ml plastic collection tubes with self-sealing caps (Kova-Tubes™) and kept frozen until analysis. As an enzyme denaturant bacteriocide and potential virus inactivator, 0.1 ml of 0.5 M $ZnSO_4$ was placed in tubes given to the hospital or clinic. This gives a final concentration of 0.005 M $ZnSO_4$ in the collected urine. Both $ZnSO_4$ and $HgCl_2$ were recently tested by Dr. Howard Faden of the Childrens Hospital for their ability to inactivate polio viruses (Mahoney polio strain). The viral solution was supplemented with one or the other of the salts and then treated with EDTA to remove excess metal ions that were themselves cytotoxic. Viral solution treated with 5 mM $HgCl_2$ did not cause infection when added to a culture of human embryonic lung tissue. On the other hand, concentrations of up to 10 millimolar $ZnSO_4$ were not effective in inactivating the virus.

Samples (5 ml aliquots) are thawed and 0.005M $ZnSO_4$ is added to samples that were collected without this "preservative". The presence of zinc was found to have a minimal but detectable effect on the molecular weight profile, so all samples that belong to ongoing series are presently processed with this reagent added. In the future $HgCl_2$ will be substituted for $ZnSO_4$ in this protocol. After reaching room temperature, an equimolar amount of EDTA is added to chelate the excess zinc ions, and the pH of the sample is adjusted to 7.2 with HCl or NaOH. Chelation of the excess zinc with EDTA, avoids

precipitation of $Zn(OH)_2$ during pH adjustments. Such precipitation could potentially alter profile patterns by co-precipitating organic constituents in the urine. A titration curve of urine indicated that pH values less than 2, near 7.2 ahd above 10 are likely to have reproducible profile patterns. In intermediate regions there appear to be partially ionized constituents having extraction yields that are highly pH dependent. In a separate series of experiments, using the Wilcoxon-WNI test to evaluate the results, it was determined that insignificant pattern changes occurred within a pH range of $\pm$ 0.5 pH unit of 7.2.

Droplets of thawed urine are also tested using Ames Multistix^{R} reagent strips to assay pH, protein, blood, glucose, ketone bodies, bilirubin and urobilinogen. These results are recorded for correlation with hospital testing of the same sample, and to indicate samples that should be excluded from data sets on the basis of abnormal kidney function.

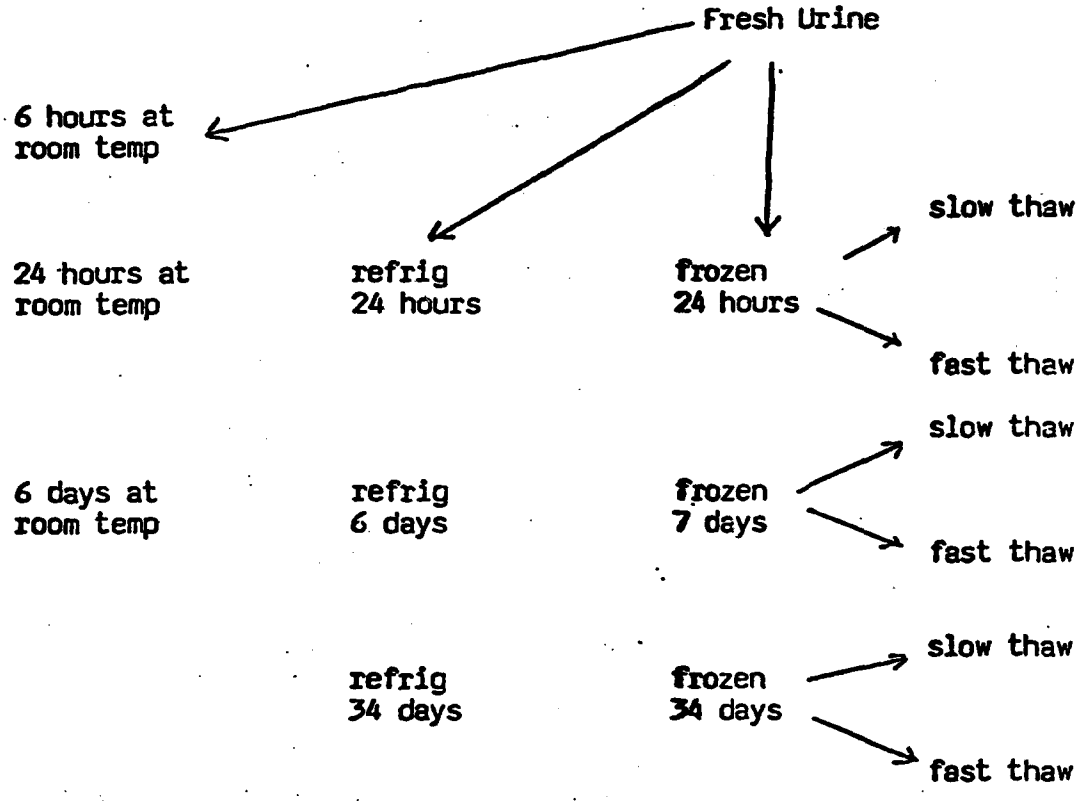
We have also initiated a study on the effects of various means of sample storage. This study is outlined in Figure 1. A volume of urine was collected from a healthy adult male. This volume was divided into 5 ml aliquots that were alternatively refrigerated, frozen and left standing at room temperature without preservative, in capped, 12 ml plastic Kova Tubes^{TM}. Three samples were prepared immediately from the fresh urine. Samples stored under the conditions outlined in Figure 1 were subsequently prepared in triplicate by the same method.

Two different methods were employed to defrost the frozen samples. At each time indicated, three tubes were allowed to defrost by standing at room temperature for approximately 45 minutes. Three other tubes were more rapidly thawed by holding them under running water. This study is still in progress and the results will be reported in next year's report.

ANBAR, MICHAEL
353-30-1453

FIGURE 1

URINE STORAGE STUDY



B. Sample Preparation Procedure

Samples are applied to the top of a 6 mm O.D. x 25 cm glass column containing 18 cm of prewashed Chromosorb P above 2.5 cm of Na_2SO_4 and a glass wool plug. The column outlet, a 10 cm length of 0.25 mm I.D. stainless steel capillary tubing, and a solvent reservoir of 12 cm O.D. x 100 cm glass tubing attached to the top of the column, are assembled using stainless steel swagelok™ unions and teflon™ ferrules (see Fig. 2). The sample is applied to the column using nitrogen pressure supplied through a swagelok fitting at the top of the column. The Na_2SO_4 serves to retain any aqueous sample that emerges from the chromosorb. After the sample is adsorbed, 5 mls of dichloromethane are added to the reservoir and forced through the Chromosorb column with N_2 pressure at a flow rate of 0.5 ml/min. The emerging eluate is continuously absorbed onto a 1 mm x 2 cm strip of Whatman GP/A glass fiber filter paper at the bottom of a conical tube. The eluate on the paper is continuously concentrated by a stream of dry N_2 directed toward the bottom of the collection tube. The dried eluate is stored in a glass vial until mass spectrometric analysis. The filter paper replaces a micro column of chromosorb previously used (see last year's report). The glass wool plugs of those sample columns frequently loosened and resulted in sample loss during loading into the mass spectrometer inlet probe. The glass fiber paper has a lower background than the chromosorb column and gives evaporation profiles similar to those obtained with the micro columns.

C. Computerized Data Acquisition

We have successfully interfaced each of our spectrometers to the FINNIGAN 2400 data system. The data system generates a digital scan function with operator selectable scan times and upper and lower mass limits. This digital scan function is applied to a 16-bit D/A converter to provide as analog signal to drive the spectrometer magnet. For use with the double focusing CID

ANBAR, MICHAEL
353-30-1453

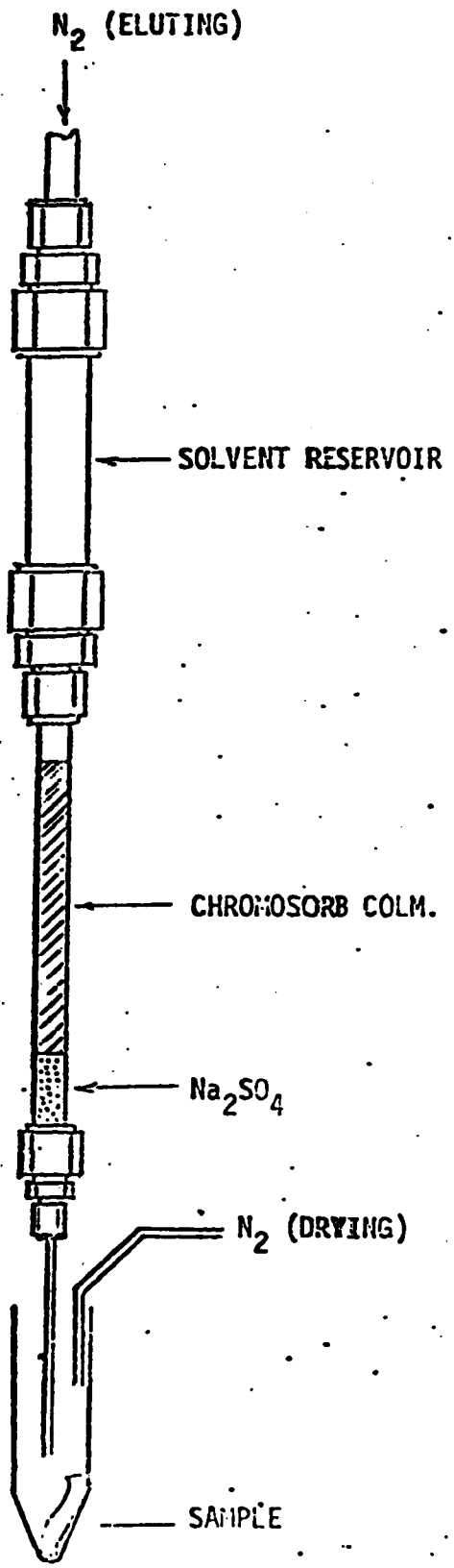


FIG. 2 SAMPLE EXTRACTION APPARATUS

instrument this analog signal is applied to a current programmable magnet power supply (Alpha Scientific Model 3048) using an Analog Device model AD284J isolation amplifier. This configuration, computer generated digital scan, D/A and current programmable supply, gives highly reproducible scans on each instrument to which the computer is interfaced. For field ionization operation, the accelerating voltage is monitored with a 4 1/2 digit digital voltmeter and maintained at the same value (nominal 5000V) at the beginning of each sample run to provide the same mass-to-time function for each analysis. The detector output is sampled by a 12 bit A/D converter with computer controlled integration times of 25 to 200 microseconds (see Fig. 3). Mass assignment is achieved by a time-to-mass calibration curve established with a known mixture of reference materials. The calibration algorithm utilizes a higher order curve fitting, capable of accurate interpolation and extrapolation for mass assignment, provided the spectrometer scan function is reproducible. In our system the digital scan function and the stable current programmed power supply gives mass assignment stability within 0.3 amu from sample to sample, including drifts due to the turning on and off the accelerating voltage supply. The magnet is continuously and repetitively scanned by the computer during the day, and acquisition of data is initiated by operator command through the data system display terminal. Mass assignment drifts of less than 1 amu are generally maintained over several days.

In spite of the inherent mass assignment reproducibility we recalibrate the instrument for each sample. Calibration for field ionization is more difficult than with electron impact. Due to the absence of fragment ions, we must calibrate with a mixture of compounds with similar vapor pressures, in order to obtain spectra with all calibration peaks present in a single scan. Presently we use a mixture of seven compounds (see Figure 4) covering a molecular weight range for m/e 73 to m/e 298. Using an instrument with a

ANBAR, MICHAEL
353-30-1453

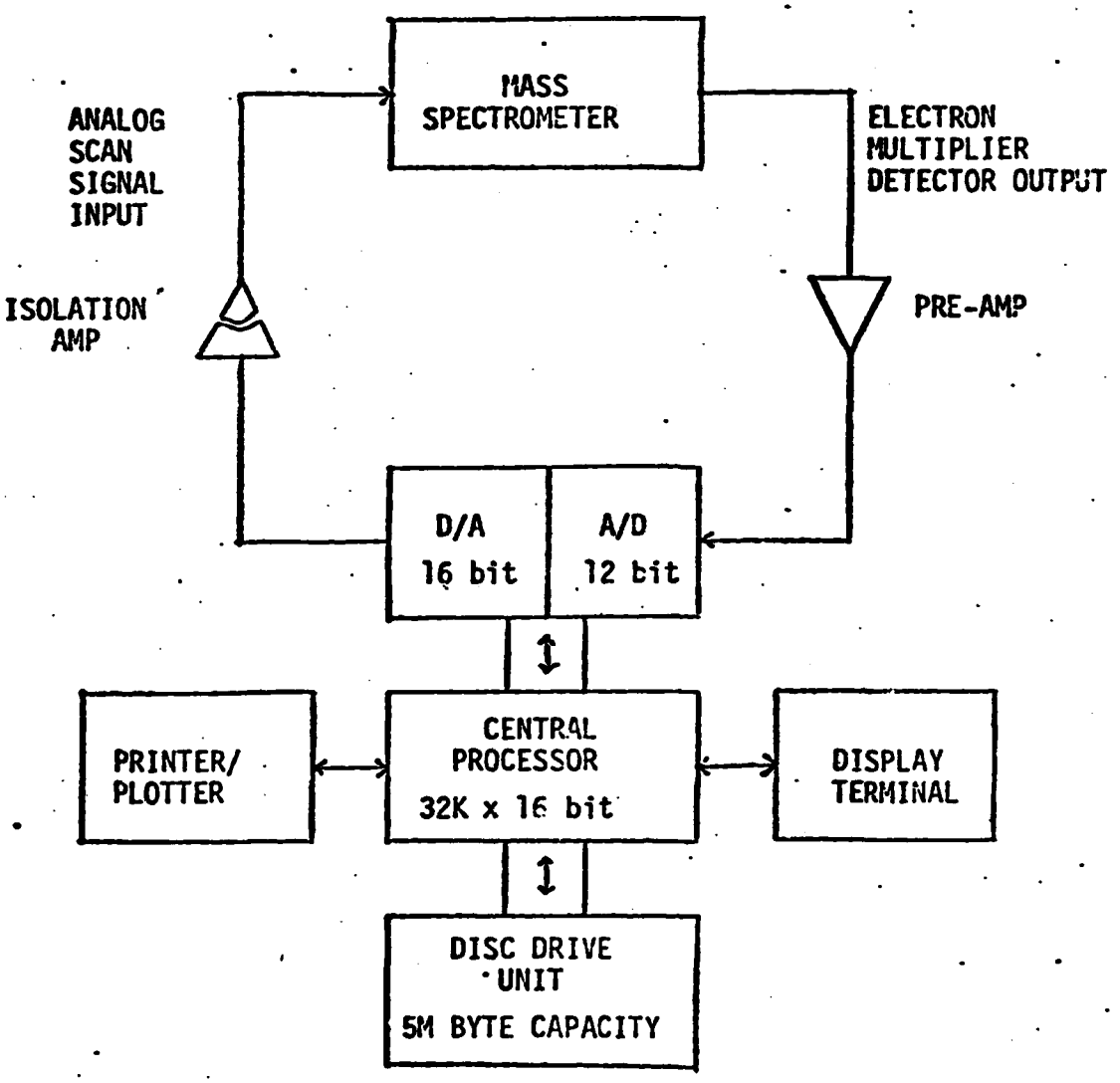


FIG. 3 MASS SPECTROMETER-DATA SYSTEM BLOCK DIAGRAM

FIGURE 4

INPUT FILE: FIC
OUTPUT FILE: FIC

MASS	INTENSITY	
73.0528	10000	DIMETHYL FORMAMIDE
108.0688	2600	ORTHO-PHENYLENEDIAMINE
152.0474	4000	PARA ANISIC ACID
170.0732	1600	PARA HYDROXYDIPHENYL
244.1253	800	TRI PHENYLMETHANE
250.1107	1000	METHAQUALONE
298.2872	400	METHYL STEARATE

MASS LIST
10/08/79 12:33:00 + 1.08
SAMPLE:

DATA: OC201 # 4 B
CAL.I: OC201 # 4 R

16 0.00 0.00 400. MINIMA MIN INTEN:
326 # 0 MAXIMA
MASS % RA % RIC INTEN.

MASS		% RA	% RIC	INTEN.
73.05	M	100.00	71.73	44992.
74.06		4.16	2.98	1870.
108.07		16.47	11.81	7409.
109.07	M	0.96	0.69	430.
152.05		4.28	3.07	1924.
170.08		3.96	2.84	1782.
244.12	M	1.99	1.43	897.
250.11	FM	2.32	1.67	1046.
298.28	M	1.55	1.11	698.

UPPER: LIST OF ACCURATE MASSES AND COMPOUNDS USED FOR CALIBRATION.

LOWER: LIST OF MASSES ASSIGNED BY CALIBRATION PROGRAM FOR SCAN TAKEN OF CALIBRATION MIXTURE.

nominal resolution of 500, the calibration with this mixture is generally accurate to 0.05 amu at mass 300. Assignment of masses outside the 73-298 amu range is by extrapolation of the computed calibration curve. Extrapolated assignment of masses up to 100 amu above m/e 298 is accurate and stable to ± 0.2 amu. This is measured by observing the assignment of a high M.W. peak (e.g. cholesterol MW 386) using several different calibration curves.

For analysis of urine extracts the mass spectrometer is scanned from 1 to 450 amu in 12 seconds. The total scan cycle time is 17 seconds including the time spent returning to the starting mass, and allowing for field stabilization. Samples are placed in the solid probe which is placed in contact with the ion source. Data acquisition starts as soon as the sample is in position near the ion source. The sample is maintained at 20°C using air cooling of the probe during sample introduction, and following initial contact with the ion source which is maintained at 200°C. The probe temperature is linearly programmed for 20°C to 200°C increase over 20 min, and held at 200°C for 5 minutes.

The data system records and stores 90 individual mass scans during the sample volatilization. These scans may be displayed in graphical or tabular format during acquisition. The data is stored in an accurate mass format with the mass of each peak assigned to within ± 60 ppm by the current calibration file. At the end of the acquisition period, the individual scans are summed following conversion of the accurate masses to nominal masses, and this integrated spectrum is used in all subsequent data processing steps. Immediately following the urine sample, a new calibration sample is introduced and 10 scans acquired over a temperature increase from room temperature to approximately 70°C. Due to the high volatility of some components in the low molecular weight range, large changes in the relative peak intensities occur during the calibration run, and only the scans which have sufficient intensity of all components of the mixture are utilized for calibration.

The drift in mass assignment of one of the components, methyl sterate, MW 298.28, is compared between this and the previous calibration file. This peak has the smallest intensity and the highest mass in the mixture, so that its mass assignment is subject to the largest variation due to ion statistics and scan irreproducibility. Differences in mass assignment for two consecutive calibration runs are typically within 0.1 amu.

After calibration the source and probe are baked out to remove any residue, in preparation for the next sample. The entire sequence of calibration, bakeout, and sample analysis requires about one hour. At least once each day the sensitivity of the ion source is measured by evaporating 1 microgram of adenine, and acquiring data with the same scan sequence used for urine samples. The integrated signal for this sample is required to be greater than 5000 ions. Correcting for the percentage of scan time actually spent monitoring the molecular ion, (10 ms/17 sec) this is equivalent to 10^{-12} Coulombs/microgram. Over the actual scan range employed this implies a minimum of 100 ions collected for 20 ng of a single component in an actual sample mixture. Examples of calibration and sample data are given in Figs. 5-9.

A standard urine extract is also analyzed at least once a week as an overall test of instrument performance. This sample is a 10 microliter aliquot of a of a concentrate obtained from the extraction of 20 ml of urine using a larger column appropriately scaled up in size to cope with a bulk sample.

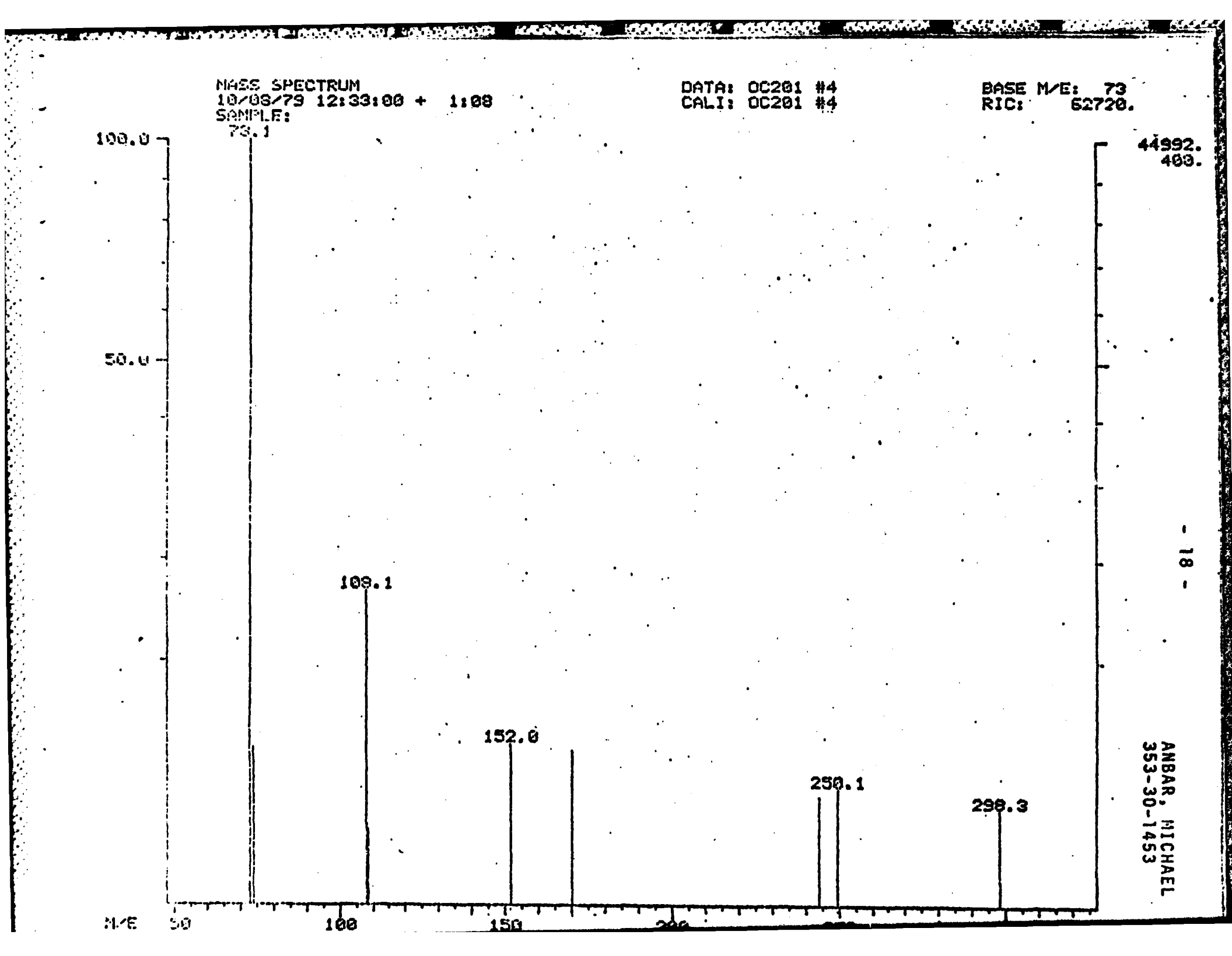
Initially, profiles were recorded using a single focusing magnetic sector spectrometer (see last year's report). In certain samples, particularly from patients with liver disorders, we encountered mass regions where ions appeared with a continuum of energies covering several amu of the mass scale. Further analysis revealed that many of these ions possessed energy-to-charge ratios

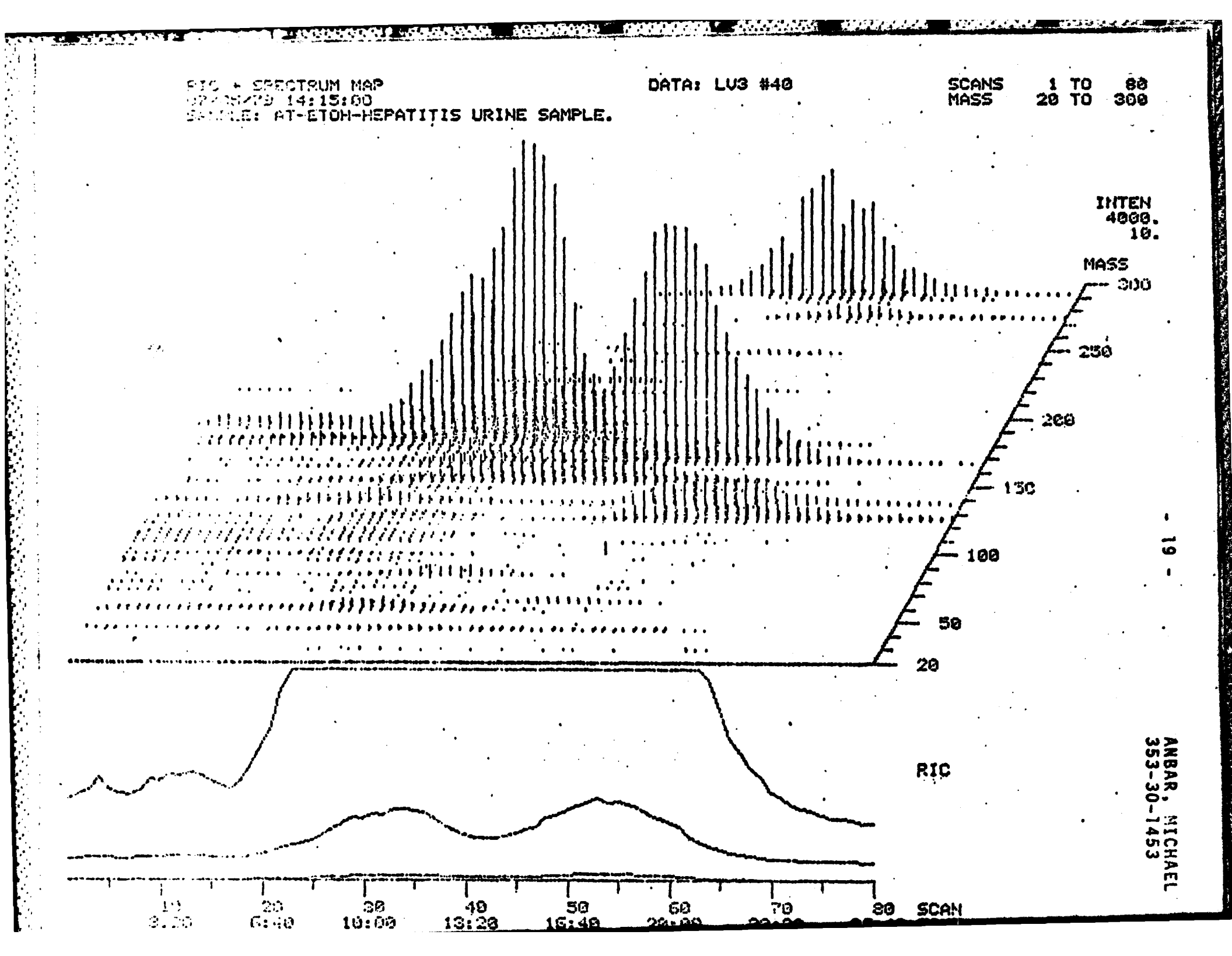
07/05/79 14:15:00 - 0.20 - 17 - PIC: 438784

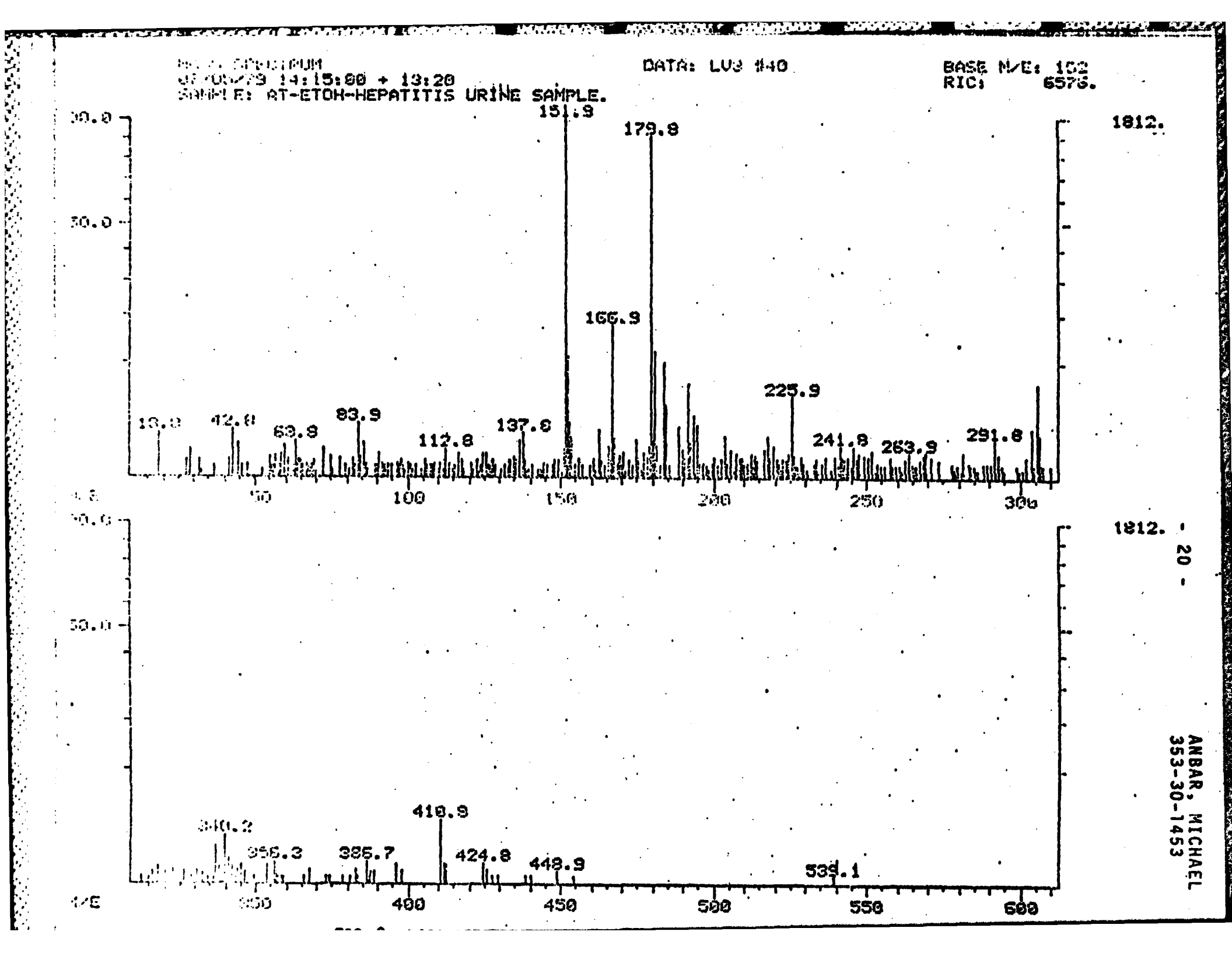
SAMPLE: AT-ETOH-HEPATITIS URINE SAMPLE

S 340 MASS	% RA	% RIC	# OF INTEN	MINIMA MAXIMA MASS	% RA	% RIC	INT
19.00?	3.97	0.41	1798.	167.00	87.32	12.32	540
29.00?	0.83	0.11	503.	168.00	9.59	1.32	58
43.00?	3.48	0.48	2108	169.00	1.55	0.21	9
45.00?	2.05	0.28	1242	170.00	0.68	0.09	4
57.00?	0.70	0.10	424.	171.00	0.60	0.08	3
59.00?	1.62	0.22	979.	175.00	0.88	0.12	5
60.00?	3.76	0.52	2276.	176.00	3.03	0.42	18
61.00?	3.52	0.49	2132.	177.00	1.18	0.16	7
64.00?	1.79	0.25	1084.	179.00	1.22	0.17	7
73.00	0.92	0.13	556.	180.00	100.00	13.80	6054
75.00	0.80	0.11	485.	181.00	16.86	2.33	1020
78.00	0.80	0.11	486	182.00	2.26	0.31	13
84.00	3.63	0.50	2200.	183.00	0.60	0.08	3
85.00	0.69	0.10	417.	184.00	20.35	2.81	1232
94.00	0.59	0.08	356	185.00	8.26	1.14	500
95.00	0.91	0.13	552	189.00	0.66	0.09	3
99.00	1.94	0.27	1176	192.00	8.56	1.18	518
100.00	1.16	0.16	700.	193.00	4.02	0.56	243
108.00	0.89	0.12	537.	194.00	43.71	6.03	2648
109.00	0.90	0.12	544.	195.00	6.43	0.89	389
110.00	1.53	0.21	926.	196.00	1.14	0.16	68
112.00	1.15	0.16	694.	208.00	1.23	0.17	74
113.00	0.72	0.10	438.	210.00	0.64	0.09	38
114.00	1.51	0.21	912.	220.00	1.13	0.16	68
117.00	0.87	0.12	528.	226.00	1.60	0.22	96
118.00	0.56	0.08	342.	248.00	1.10	0.15	66
122.00	0.91	0.13	553.	274.00	4.96	0.68	300
123.00	0.57	0.08	346.	275.00	1.15	0.16	69
124.00	1.95	0.27	1180.	288.00	0.89	0.12	53
125.00	18.50	2.55	11200.	290.00	0.69	0.09	41
126.00	7.78	1.07	4712	291.00	0.99	0.14	60
127.00	0.89	0.12	535.	292.00	43.45	5.99	26304
128.00	1.36	0.19	821.	293.00	9.00	1.24	5448
129.00	0.75	0.10	455.	294.00	0.63	0.09	382
131.00	0.58	0.08	349.	301.00?	0.58	0.08	352
135.00	0.61	0.08	372.	302.00?	11.58	1.60	7008
137.00	14.24	1.97	8624.	303.00?	2.68	0.37	1624
138.00	8.09	1.12	4896.	304.00?	42.60	5.88	25792
139.00	1.74	0.24	1054.	305.00?	7.21	1.00	4369
146.00	0.81	0.11	493.	306.00?	43.76	6.04	26496
147.00	0.61	0.08	370.	307.00?	6.83	0.94	4134
148.00	0.57	0.08	345.	340.00?	5.93	0.83	3520
151.00	0.90	0.12	542.	341.00?	2.43	0.34	1470
158.00	42.66	5.91	25958	356.00?	1.28	0.18	777
159.00	7.70	1.06	4684	357.00?	0.95	0.13	577
164.00	3.13	0.43	1904	358.00?	1.32	0.18	803
169.00	0.67	0.09	408	359.00?	0.75	0.10	456
170.00	0.55	0.08	334	367.00?	0.71	0.10	432
176.00	4.78	0.66	2911	411.00?	1.38	0.24	836
177.00	0.71	0.09	433	412.00?	6.49	0.89	3951
186.00	3.01	0.18	795.	413.00?	1.84	0.25	1119

FIGURE 5


MASS SPECTRUM
10/08/79 12:33:00 + 1:08
SAMPLE:
DATA: OC201 #4
CALI: OC201 #4
BASE M/E: 73
RIC: 62720.
73.1
100.0
50.0
44992.
400.
109.1
152.0
250.1
298.3
M/E
50
100
150
- 18 -
ANBAR, MICHAEL
353-30-1453


RIC + SPECTRUM MAP
DATA: LV3 #40
SCANS 1 TO 80
MASS 20 TO 300
SAMPLE: AT-ETOH-HEPATITIS URINE SAMPLE.
INTEN
4000.
10.
MASS
300
250
200
150
100
50
20
RIC
SCAN
ANBAR, MICHAEL
353-30-1453
- 19 -


DATA: LU3 #40
BASE M/E: 152
RIC: 6576.
SAMPLE: AT-ETOH-HEPATITIS URINE SAMPLE.
1812.
151.9
179.8
166.9
225.9
83.9
137.8
112.8
63.9
42.8
241.8
263.9
291.8
410.9
356.3
385.7
424.8
448.9
539.1
M/E
1812.
ANBAR, MICHAEL
353-30-1453

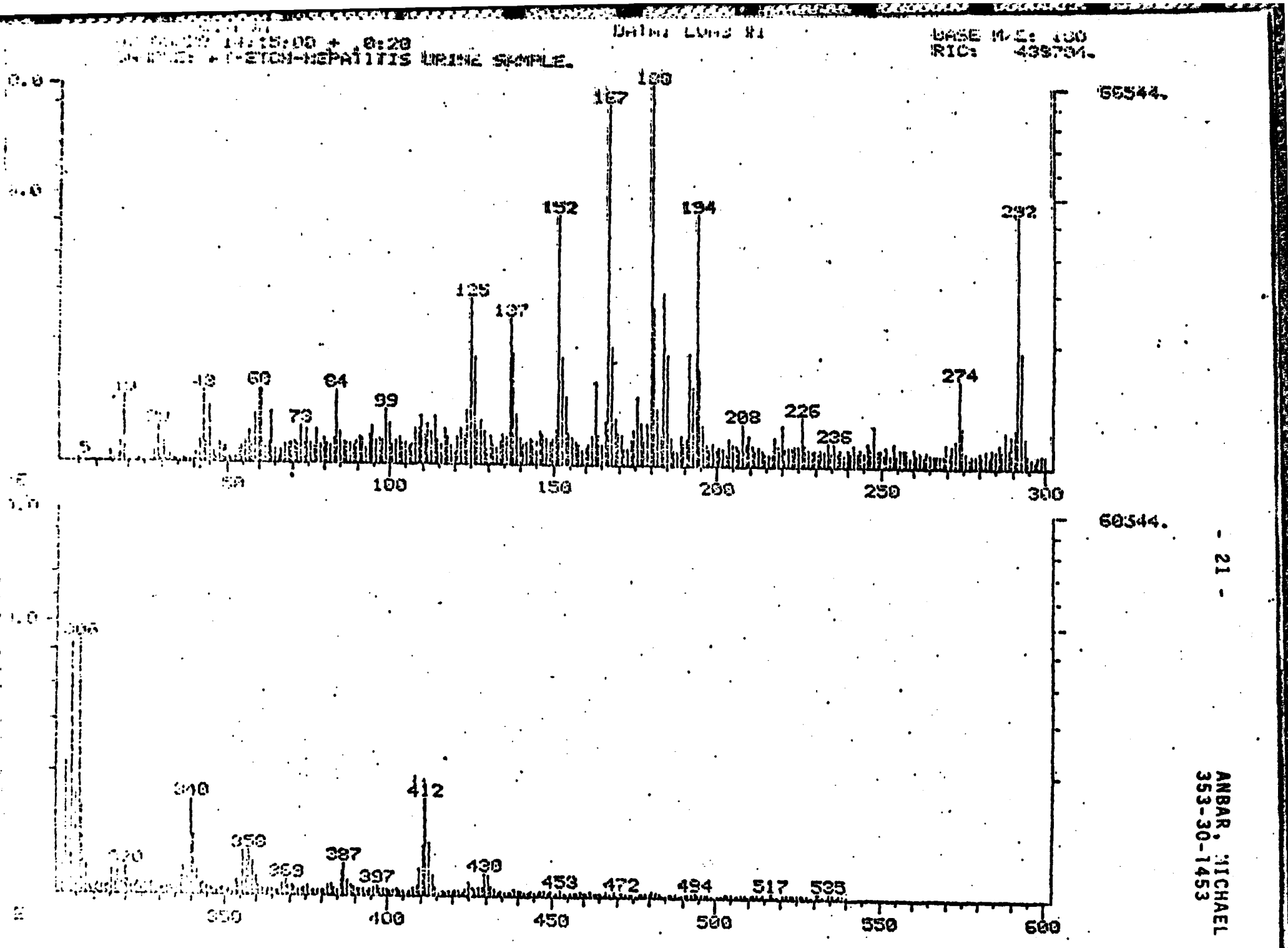


FIG. 9 INTEGRATED SPECTRA OF SAMPLE FROM ALCOHOLIC LIVER DISEASE PATIENT

ANBAR, MICHAEL
353-30-1453

greater than the main beam energy. These ions are a result of the formation of doubly charged species, which upon collision with neutral molecules either gain an electron or decompose to smaller fragments. In our instrument the most likely place for such collisions to occur is in the lens between the source and the object slit. This region has the highest density of neutral molecules as well as of focused deccelerated ions. The distribution of energies arises from the fact that these processes may occur anywhere between the ionizer and the object slit in a region with a potential gradient. The final energy of the ion will be a function of the position where the transition from doubly charged to singly charged species occurred. The acceleration of the ion as a doubly charged species, even over a small fraction of the net accelerating voltage, will result in an ion with a final energy greater than the main beam energy. In order to obtain unit mass resolution of the singly charged normal molecular ions, a double focussing instrument employing an energy analyzer is required. Since the extent of these processes is variable and may occur in any sample, we now analyze all samples on the reverse geometry double focussing instrument (the CID instrument) described in our original proposal.

D. Data Reduction and Transmission

The mass spectra acquired are processed locally to facilitate transmission to the University's main computer (Control Data Corporation CYBER 173). A second set of operations in the CYBER then follows, and results in spectral data suitable for the subsequent statistical operations.

Within our dedicated Finnigan/INCOS computer system the individual mass spectrometer scans are stored as acquired. Upon completion of the run a single spectrum is produced by addition of all scans. This spectrum is then translated into a fortran-readable format suitable for transmission to the CYBER main computer over a 1200 bund multiplexor phone link.

When the spectra are on file in the CYBER, format and simple logical checks are made on the data, which are then corrected as required. Each spectrum is then normalized to unit total area, with individual peak areas greater than 5 per cent of the total excluded from the normalization sum)fpr tje ratopma;e see last year's report). All statistical analysis is then accomplished using these normalized spectra.

3. DIAGNOSTIC STATISTICAL ANALYSIS.

A. Introduction

During 1979 we have devoted a substantial effort to evaluate our diagnostic statistical analysis comparing a number of alternative techniques for the selection of variables and for the separation of cases into diagnostically meaningful groups.

The statistical analysis in this project has the following objectives:

a. To separate the biological samples (cases) into statistically distinct groups correlated with the disorder of interest (e.g. healthy vs. diseases, healthy vs. alcoholic liver disease, pneumonia vs. bronchitis, bacterial vs. viral pneumonia, etc.) by a characteristic set of variables.

b. To assign correctly an unknown case to one of a number of pre-specified groups, which have been previously developed on the basis of a continuously increasing learning set.

c. To identify the variables that best characterize a given pathological state, in order to facilitate the understanding of the biochemical nature of the disease and possibly also in order to explore the possibility of quantitative assay of the particular metabolite by a simpler non-mass spectrometric analytical technique.

d. To identify variables that co-vary, for two reasons - first to attain a better understanding of the biochemical nature of the disease, and second to minimize undue weighting bias in the differentiation of cases into diagnostic groups.

Different multivariate analysis techniques have provided us with answers to these questions with different degrees of efficiency. As stated above we have started to compare different statistical techniques for their merits in meeting our objectives. In the following sections we shall discuss the current status of this evaluation, which will be continued during the forthcoming stages of this project.

We may separate our current statistical analysis into two phases - the selection and rating of variables (mass spectral peaks) according to their diagnostic value, and the classification of cases (patient' urine or tissue culture media) into groups.

B. Selection and Rating of Variables

Each normalized mass spectrum comprises hundreds of variables (peaks) each of which represents the concentration of a metabolite (or a group of metabolites sharing the same nominal mass) in the biological sample. When we compare the magnitudes of a given variable in samples coming from two biochemically distinct groups we observe three types of variation:

- variation due to the analytical procedure
- variation due to biological variance (due to genetic or nutritional factors), and
- variation associated with the experimental difference between the groups, e.g. variation due to the pathological status of a human subject, or the infected status of a tissue culture.

Ideally there should be no variation of the first kind and the variation due to the pathological status should be by far larger than the biological variation. In reality, however, the majority of variables show a large biological variation (on top of a finite experimental variance) and are thus of minimal diagnostic value (we shall call these diagnostically useless variables). Our problem is, therefore, to select those variables which may be diagnostically useful.

Any statistical diagnostic procedure will become ineffective if given an excessive number of diagnostically useless variables, even in the presence of useful ones. On the other hand, we would like to utilize every useful variable since each of these increase the diagnostic power of the classification procedure. An acceptable variable selecting procedure has, therefore, to identify and reject useless variables, while retaining all the useful ones. Moreover since there will always be variables more "useful" than others, an adequate procedure should rate them accordingly, and thus allow us to optimize the diagnostic power of the diagnostic procedure.

There is another factor that should be taken into account - covariance. In a biological system there are many variables that are biochemically interrelated, so that their variation associated with a given pathological state is interdependent. If such variables are used in a statistical method based on a pattern of a given number of independent variables, they may bias the result by giving an undue high weight to a single variation (accompanied by a set of dependent variables). It would be advantageous, therefore, to identify such co-variances and eliminate the satellite variables from the diagnostic classification pattern. A desirable feature of a selecting technique would be, therefore, the ability to identify covariance and minimize its effect.

B1. The Wilcoxon Test

This non-parametric ranking of variables according to the probability of being constituents of the same population has been described in our previous reports. This ranking has two important shortcomings: first, it does not identify covariance and second, it will not discriminate against artifactual deviants. In fact this treatment may give a variable (peak) with an unlikely large deviation in one of the cases (spectra) an undue outstandingly low probability. This second shortcoming is avoided in the t-test variable rating.

B2. The t-test rating

We have applied the t-test program P3D of the BMDP programs package (UCLA 1977) to determine the null hypothesis that each of the variables in two tested groups of cases belongs to the same population. Unlike in the Wilcoxon test this probability is calculated on the basis of deviations from the group's average, while taking into account the difference in values between the two group averages. This program also provides us with the variance of each variable in each group, so that artifactual deviants can be readily identified and discounted.

In spite of the intrinsic differences between the Wilcoxon and the t-test, the two programs identified 50 out of 400 peaks from the same test set of spectra as of prime diagnostic value with an overlap of over 90% of the variables selected, (although the order of ranking by the two procedure was somewhat different). In view of this finding and since the t-test provides additional useful information we prefer now to use this program for selection of the diagnostic peaks.

B3. The Stepwise Discriminant Analysis Procedure.

This procedure to be discussed below, selects and rank orders variables according to their "F" values. The F value for each variable is proportional to the square of the intergroup difference and inversely proportional to the square of the intragroup variance around that group's average. Since this procedure requires considerably more computer capacity than the two preceding methods, it can handle just a limited number of variables (50 peaks in our case). This limitation requires pre-selection of variables by one of the preceding techniques, to be followed by their F value ranking according to their usefulness in separating the cases into distinct groups. The main feature of this statistical procedure is its ability to identify and reject co-variants. In spite of this important advantage the discriminant analysis can hardly be considered a practical peak selecting procedure, because of its high demand on computer time and capacity. However since this technique is being used as a group classification and case assignment procedure, the peak selection and ranking according to the F values may be considered as a fringe benefit.

B4. Modes of use of selected variables

The variables for a given classification procedure can be selected by virtue of meeting a certain arbitrary criterion (e.g. having a p value below a given value), or by rank ordering according to a given criterion (e.g. starting with the variable of lowest p value, followed by the next lowest, and so on) and then picking an arbitrary number (e.g. 50) having the lowest p values. The selected variables can then be used in the group classification procedures without any weighting.

Alternatively a characteristic classification parameter (e.g. the p value) can be used as a weighting factor. Since the diagnostic usefulness increases as p decreases, using 1/p as a weighting factor is perhaps the simplest weighting procedure. This weighting will, however, give variables with very small p values a very high weight. Alternative weighting factors could be for instance $1/\sqrt{p}$, $1/\sqrt[3]{p}$, or log 1/p, which would decrease the overweighting of variables with very small p values.

There are advantages to either peak selection procedure. The discriminant method (by an arbitrary cut-off) requires human judgement for each set of cases. This shortcoming can be eliminated by rank order cut-off. Using the t-test one can use the variance in addition to the t or p values as a second criterion in selection.

The weighting procedure while free from subjective intervention nevertheless requires optimization to obtain the best use of separating variables. However, the weighting of variables as in the WNI procedure (see our 1978/9 report) involves the choice of an arbitrary weighting function which is at best a compromise between an optimized use of the variables and a use based on a subjective threshold.

C. Group Classification Procedures

During the previous phase of this program we have used basically just one classification procedure, namely the weighted non-parametric index (WNI) method which has been described in our previous reports.

This procedure which is simple and straight forward has certain limitations. First, it is applicable to only two groups of variables. Although when WNI (1) is plotted vs. WNI (2) one can obtain some secondary clustering, indicating sub-groupings, but the separation between these is always by the original choice of the groups, since the average value of each variable in each group is the reference point for the WNI calculation. Second when small sets of cases are analyzed, the WNI values are strongly influenced

by variables with large variances from their respective group averages. This is especially true when the difference between the group average is relatively small. Third, the diagnostic referent point on the D = WNI (1) - WNI (2) scale is arbitrary, which becomes problematic if the D values for members of the two groups form a progressive continuum without a significant gap between D's of the two groups. Fourth, this "non-parametric" procedure does not provide us with a measure of probability of a given case belonging to each group, thus it is lacking a quantitative measure for the diagnostic assignment of a given case to a particular group.

In view of these limitations, we have experimented with two other classification procedures - the clustering analysis procedure (P2M procedure BMDP UCLA, 1977) and the stepwise discriminant analysis procedure (P7M BMDP UCLA, 1977). The former classification is free of the bias of assignment of cases to a particular number of groups, whereas the latter can handle efficiently a large number of pre-specificed groups and it provides the probability of each case belonging to any of the groups in question.

C1. The Clustering Analysis

In this procedure one represents each of the cases (each with n variables) as a point on a surface of a n dimensional space. This is done by calculating a n dimensional vector as a resultant of the values of the variables measured on n orthogonal coordinates. If we have many cases, each constituting a point on the n dimensional surface, we can calculate the distance on this surface between any given two points. The clustering procedure selects the two points with the shortest n dimensional Euclidian distance between them, producing a cluster of two. Then the program tries to find among all the remaining points a (third) point (case) closest to the first two points, forming a cluster of 3. Again the program selects and registers a (fourth) point closest to the

cluster of three, and so on. When the distance between the growing cluster and the next point is larger than between a pair of the remaining points, a new cluster of two is selected, which can again grow, aggregating points in its vicinity. This process continues until the nearest distance remaining is the distance between the boundaries of two clusters, which are then registered as a cluster of clusters. The classification is ended when all points are accounted for, when a master cluster containing all the points (all the cases) is registered. This approach is completely bias-free as far as the number of groups (clusters) it will form from the set of cases; only after clustering can one check the a priori assignment of a given case against the cluster it ended up in. The program also allows the identification of the relative positions of points (cases) within clusters. On the other hand the procedure does not test the variables for their variances from a group average (which is done in the WNI and in the discriminant analysis classifications) or for co-variance, which is performed in the discriminant analysis.

C2. The Stepwise Discriminant Analysis

This classification procedure separates the cases into a prespecified number of groups after analyzing the variance of each variable. This procedure also selects those variables which separate the cases into the specified groups most effectively.

The procedure first determines the variance of each variable within each group and compares it to the variances between groups. The comparison is done by calculating F values, i.e. dividing the square of intergroup variance (S_G) by the square of the variances (S_v) of the individual variables around the corresponding group average: $F = S_G^2/S_v^2$. The program then selects the variable with the highest F value and if this is larger than a pre-specified threshold "F-to enter" it will use this variable to classify the cases into groups.

In the second step it will test all remaining variables for their ability to separate between groups on a 2 dimensional surface, comparing again the new intergroup variance to the group variances. The variable with the highest "F to enter" value at this step will be then added to the first selected variable, provided its F value exceeds the "F to enter" threshold and provided that a correlation coefficient between the two variables is not above a specified limit. A high correlation would obviously invalidate the group classification by two presumably independent variables.

After the two classifying variables are selected the program computes for each case a point (vector) on a 3 dimensional surface using each of the remaining variables combined with the 2 variables selected in steps 1 and 2. The variable with the highest F value is selected, provided it exceeds the threshold and that it does not co-vary with the combination of the two first variables. If it does not fulfill the second condition the variable with the next highest F value is selected and tested against the two criteria. The variable selection procedure is continued until all remaining variables end up with F values below the threshold or exhibit excessive covariance with the set of selected variables.

Following each step the program also re-checks the F values of all the variables selected for classification up to that step, since their F values will change with each added variable. If any of the previously selected variables has an F value lower than a given threshold ("F to remove") it will be removed from the classification set, and the procedure is repeated again for all the remaining non-classifying variables to select a new one with an acceptable F value and covariance coefficient. The variables selected at each step are combined to form a linear, optimized classification function (n-dimensional vector that maximizes the [illegible] ratio.

Once the variable selection process is complete using say m variables, the cases are classsified into groups as points on a m dimensional surface and the intercase distance is calculated. From these the centroids or group averages are computed as is the probability of assignment of each case to any given group. The grouping can be graphically presented two-dimensionally as points indicating the groups centroids or as clusters of points (representing the individual cases) around the centroids.

D. A Comparative Evaluation of the Statistical Classification Procedures.

As stated above each of the 3 classification procedures tested by us has more desirable and less desirable features when compared to the others. To illustrate this we applied the 3 procedures to the same set of data, namely analysis 39 spectra from urine of 14 patients with alcoholic liver disease compared with 26 spectra from 13 healthy adults.

The results of the Wilcoxon test on 100 peaks of lowest p values is given in Figure 10. The WNI values of the 65 cases for the two groups using 1/p weighting as well as the values of D = WNI (1) - WNI (2) are given in the same figure. One can see here that all the D values of the pathological samples (LV) are negative and smaller than -35 whereas those of all the controls (CP) are positive, with the exception of case CP22 which is negative (-9.5) but still significantly larger than any of the LV's. In other words this classification did not show any overlap between the tested groups. A computerized graphical presentation of the WNI data is presented in Figure 11 where each case is presented by its WNI (1) and WNI (2) values. We see here the clustering of the two groups with a distinct region of demarcation between them. Although WNI (1) = WNI (2) for a case would indicate that it equally belongs to the two groups this is not necessarily true when applied statistically to two groups where at least one cluster has D values very different from zero.

FIGURE 10

161	.00000031	116	.00000085	231	.00006948	192	.00065667	245	.00326195	170	.00884791
158	.00000031	146	.00000176	81	.00007691	18	.00065667	186	.00348539	214	.00993402
197	.00000002	168	.00000076	136	.00008504	200	.00071093	122	.00372204	241	.01351855
330	.00000015	160	.00000104	118	.00013814	202	.00073160	147	.00397270	201	.01051855
85	.00000067	239	.00000135	145	.00015174	148	.00069852	206	.00397273	149	.01051855
79	.00000012	112	.00000135	188	.00015174	152	.00069852	285	.00451879	247	.01051855
150	.00000018	234	.00000614	175	.00015997	265	.00112901	153	.00461546	216	.01351855
121	.00000018	194	.00000713	142	.00021884	141	.00112901	77	.00512908	226	.01113820
140	.00000023	86	.00000949	211	.00023920	31	.00151755	286	.00546034	181	.01245139
224	.00000028	169	.00001068	318	.00026140	163	.00151755	230	.00546034	227	.01245139
208	.00000030	203	.00001416	138	.00026140	165	.00163155	54	.00546034	71	.01393697
337	.00000032	340	.00001608	139	.00031121	105	.00163155	129	.00581003	287	.01467770
117	.00000034	78	.00002057	218	.00040184	240	.00175537	61	.00617901	229	.01549071
96	.00000042	207	.00002057	319	.00043691	29	.00188255	350	.00617901	215	.01534130
94	.00000044	219	.00002319	137	.00043691	178	.00232323	198	.00597827	179	.01634130
104	.00000050	43	.00002608	159	.00051535	73	.00240913	97	.00741037	238	.01634130
164	.00000050	139	.00003669	228	.00051535	246	.00285258	251	.00786536	217	.01634130
198	.00000051	60	.00004566	250	.00055910	187	.00305123	88	.00834420	101	.01634130
102	.00000054	193	.00005082	74	.00060614	193	.00335123	233	.00884791	303	.01723077
102	.00000055	205	.00005082	225	.00060614	190	.00326195	213	.00864791	242	.01816045

TAG	WVI(P)	WVI(2)	D
LV1	22.2279	61.8611	-39.6332
LV2	19.8587	59.4314	-39.5627
LV3	16.7556	70.1935	-53.4249
LV4	31.2241	77.4078	-46.1838
LV5	29.8909	79.4271	-49.5361
LV6	36.5374	80.3291	-43.7917
LV7	29.9554	79.3170	-49.3615
LV8	21.3087	59.6871	-38.3783
LV9	17.7453	64.2972	-46.5518
LV10(X5)	26.3268	62.7830	-36.7561
LV11(X6)	21.7947	63.7182	-41.9235
LV12(X1)	12.6833	69.4226	-56.7193
LV13(X2)	16.4550	71.2143	-54.7593
LV20	8.8986	70.2593	-61.3607
LV21	18.4758	63.9214	-45.4456
LV22	10.0309	68.4761	-58.4362
LV23	11.3226	72.3791	-61.0556
LV24	20.2200	78.8499	-59.6298
LV25	23.4524	76.9839	-53.5315
LV26	19.1874	59.9524	-40.7150
LV27	13.8383	65.2102	-51.3719
LV28	15.4991	65.7579	-50.2588
LV29	11.7760	67.0595	-55.2835
CP0	52.9444	28.4967	24.4478
CP1	61.3937	17.4223	43.9715
CP2	57.0847	37.8311	19.2536
CP3	51.9014	33.0917	18.8097
CP4	54.5306	29.2925	25.2441
CP5	54.6431	25.0456	29.5976
CP6	54.0397	30.2827	23.7270
CP7	68.1972	12.8667	55.2406
CP8	49.8786	37.7411	12.1375
CP9	61.3780	40.4931	20.8849
CP10	79.9826	21.5971	49.3855
CP11	59.6015	10.8920	58.7094
CP12	51.2927	33.2505	18.0422
CP13	52.3912	30.4874	21.9038
CP14	56.7267	21.7985	34.9383
CP15	55.6018	23.1143	32.4875
CP16	51.1059	29.3202	21.7858
CP19	47.4577	35.9974	11.9603
CP20	49.4696	33.4420	16.0277
CP21	55.9513	28.9414	27.0139
CP22	37.5939	47.0697	-9.4758
CP23	54.6401	19.1939	45.4462
CP24	53.3539	34.9375	28.4135
CP25	72.6553	46.5704	26.0848
CP26	73.5459	24.6155	48.9304

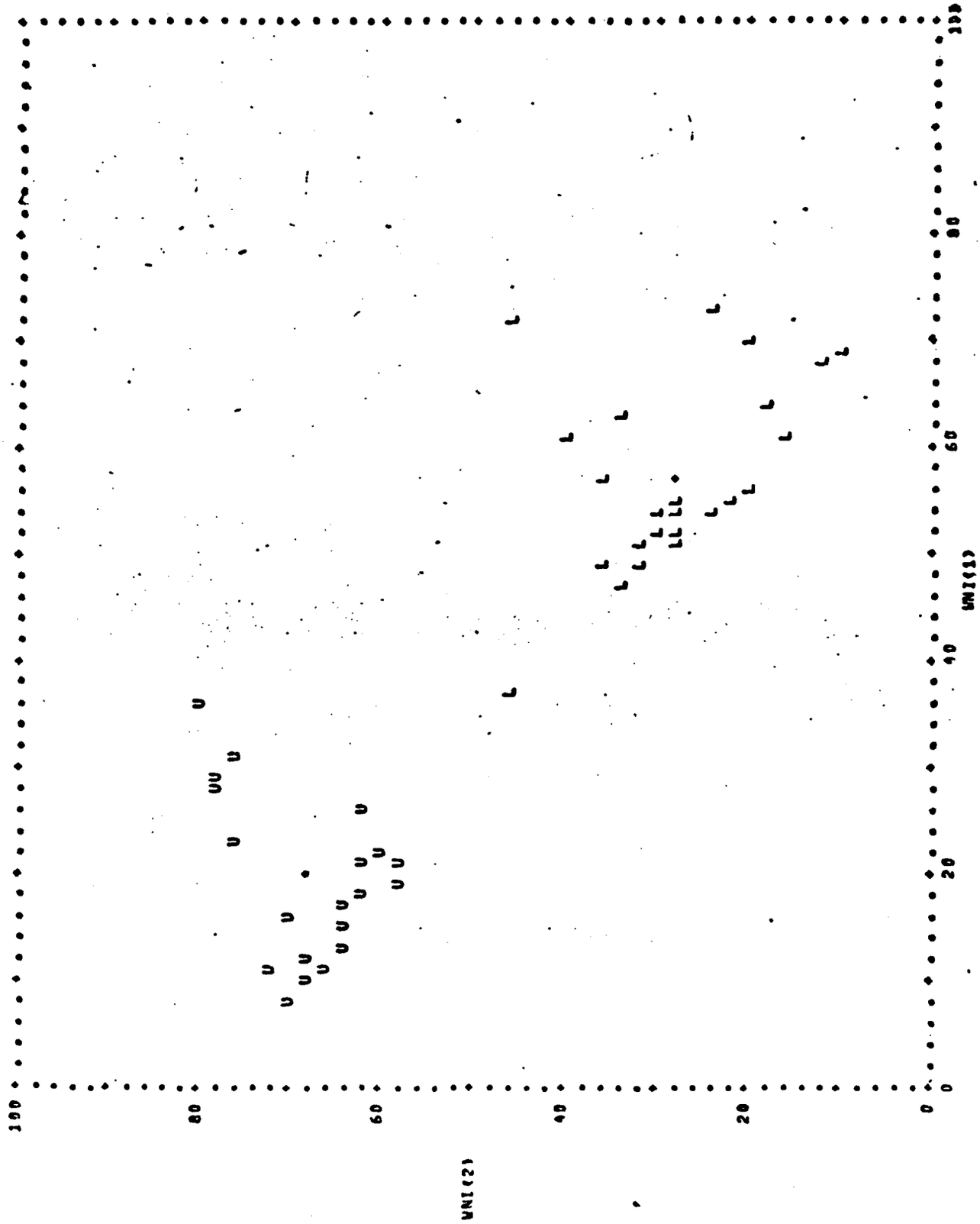


FIGURE 11

The clustering analysis applied to the same cases presented in Figure 12. Here cases 1 to 23 are of liver patients, and cases 24-48 were of controls. We see here that clustering began with case 17, 13 and 6 followed by cases 19 and 11. By the end of the clustering process all samples 1 to 24 (the pathological cases) were in one cluster with case 44 (of the controls) being the next one added to this cluster (but only in step #41 in the amalgamation order). All the other controls are again in a distinct different cluster. One may also distinguish some sub-clusters like cases 1, 2, 10 and 11 or 15, 21, 17, 16, and 20 among the pathological samples. Since this program does not presume any predetermined groups for classification these subgroups may have some biochemical features in common in addition to being part of a liver disease population.

Figure 13 presents part of the computer output of the Euclidian distances between the points representing each case on an n dimensional surface. (The distance between 48 points to just 15 points of other cases are shown in this figure). These distances are then amalgamated or clustered as discussed above.

Figure 14 presents the same variables and their respective intragroup averages used for the subsequent stepwise discriminant analysis.

Figure 15 shows the first two steps of selection of the two first variables for classification, namely mass peaks 197 and 95 respectively. The same figure then describes step #15 of the procedure by which 15 peaks were selected by the "F-to-enter" and covariance criteria.

Figure 16 presents a later stage in the procedure - step #40, when just 30 variable (peaks) were selected as parameters in the classification, incicating that 10 variables originally selected were subsequently rejected due to F values below the ("F-to-remove") threshold.

ANBAR, MICHAEL
353-30-1453

FIGURE 12

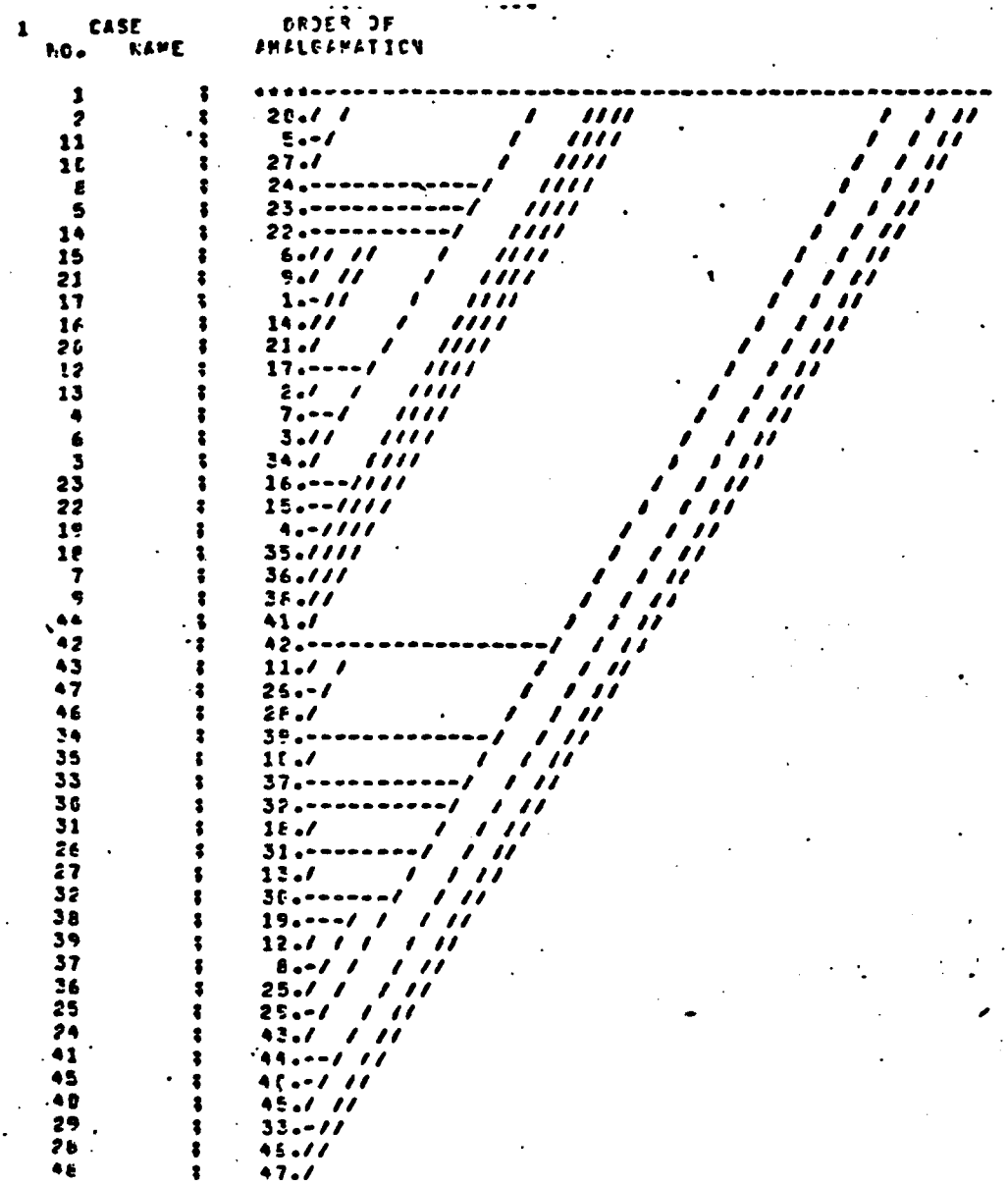


FIGURE 13

INITIAL DISTANCES BETWEEN CASES

CASE NUMBER	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.00	4.92	8.51	9.51	7.28	10.43	9.49	8.82	9.58	5.76	6.22	6.90	7.62	8.84	9.71
2	4.92	0.00	8.11	9.21	6.78	10.04	8.71	7.44	9.58	5.81	6.25	6.81	7.48	7.99	8.59
3	8.51	8.11	0.00	2.78	6.26	4.18	6.95	5.46	7.06	7.72	5.89	4.30	4.41	5.46	6.83
4	9.51	9.21	2.78	0.00	7.18	2.35	7.08	6.86	7.61	8.26	6.50	4.76	4.66	5.97	7.64
5	7.28	6.78	6.26	7.18	0.00	7.84	5.36	6.05	8.17	6.35	5.90	5.27	5.45	5.94	6.34
6	10.43	10.04	4.18	2.35	7.84	0.00	7.67	8.14	8.22	8.88	7.16	5.69	5.58	6.48	8.00
7	9.49	8.71	6.95	7.08	5.36	7.67	0.00	8.24	9.09	8.17	7.53	6.58	6.69	6.93	8.00
8	8.82	7.44	5.46	6.86	6.05	8.14	8.24	0.00	8.65	8.56	7.72	6.27	5.83	5.88	5.93
9	9.58	9.58	7.06	7.61	8.17	8.22	9.09	8.65	0.00	9.38	8.36	7.36	6.47	5.49	5.93
10	5.76	5.81	7.72	8.26	6.35	8.88	8.17	8.56	9.38	0.00	3.13	5.70	6.59	7.98	9.16
11	6.22	6.25	5.89	6.50	5.90	7.16	7.53	7.72	8.36	3.13	0.00	4.25	5.10	6.82	8.33
12	6.90	6.81	4.30	4.76	5.27	5.69	6.58	6.27	7.36	5.70	4.25	0.00	2.01	4.68	7.02
13	7.62	7.48	4.41	4.66	5.45	5.58	6.69	6.93	6.47	6.59	5.10	2.01	0.00	4.35	6.49
14	8.84	7.99	5.46	5.97	5.94	6.48	6.93	5.88	5.49	7.98	6.82	4.68	4.35	0.00	3.29
15	9.71	8.59	6.83	7.64	6.34	8.00	8.00	5.93	5.93	9.16	8.33	7.02	6.49	3.29	0.00
16	8.75	7.91	6.72	7.37	6.24	7.76	7.69	6.95	6.89	8.77	7.73	6.03	5.71	4.72	4.84
17	8.50	8.01	6.49	6.89	6.03	7.08	7.44	7.13	6.83	8.49	7.40	5.70	5.31	4.66	5.02
18	9.42	7.93	7.72	8.13	6.64	8.54	6.95	8.32	9.73	7.27	7.37	7.44	7.87	7.32	7.89
19	9.24	7.85	7.36	8.01	6.27	8.43	6.28	7.78	9.31	7.65	7.39	7.15	7.60	6.65	7.21
20	7.69	6.85	6.37	7.39	6.44	8.01	7.88	6.80	8.76	8.01	7.10	6.55	7.01	5.99	6.11
21	8.39	7.55	6.25	7.02	6.24	7.52	7.27	6.32	7.29	8.36	7.45	6.24	6.02	3.97	3.68
22	10.82	9.10	8.91	9.38	8.45	9.52	8.49	9.29	10.99	9.29	8.97	8.77	9.10	7.74	8.11
23	9.99	7.43	8.01	8.96	7.04	9.53	7.99	8.26	10.06	8.20	7.98	7.58	8.09	6.75	7.02
24	12.08	8.64	8.79	10.32	8.76	11.28	10.75	7.12	10.72	10.81	10.28	9.43	9.56	9.42	9.29
25	11.86	10.35	8.93	10.32	9.90	11.18	12.05	7.97	11.59	12.05	11.22	10.10	10.40	9.94	9.39
26	12.28	10.75	10.47	11.91	11.09	12.77	12.89	9.12	11.47	12.78	11.90	10.83	10.90	9.64	9.43
27	10.76	9.21	8.26	9.77	8.88	10.68	10.53	7.04	9.53	11.06	10.09	8.68	8.64	8.01	8.28
28	14.76	13.24	12.56	14.42	13.32	15.29	14.89	10.84	14.43	15.05	14.38	13.54	14.04	13.91	13.42
29	13.60	12.10	11.59	13.30	12.07	14.08	13.95	10.13	13.54	13.63	13.07	12.58	12.86	12.76	12.47
30	13.66	11.17	10.21	11.92	11.55	13.21	13.23	8.77	11.86	13.51	12.64	11.39	11.47	11.03	10.64
31	12.34	10.33	9.72	11.11	10.49	12.26	12.17	8.23	10.77	12.37	11.42	10.15	10.20	9.22	9.19
32	11.49	9.36	9.27	10.82	10.24	11.79	11.90	7.65	10.64	11.68	10.58	10.12	10.19	9.85	9.92
33	13.52	11.43	11.51	12.82	12.85	13.62	14.07	8.85	12.74	13.70	12.78	12.30	12.34	11.21	11.11
34	13.01	10.78	10.33	11.75	11.40	12.89	12.71	9.18	11.59	12.88	11.93	10.81	10.94	9.91	10.02
35	13.72	11.13	10.89	11.80	12.04	13.00	12.98	9.76	12.04	13.30	12.31	11.32	11.38	10.60	10.97
36	11.30	9.29	8.17	10.96	10.03	11.59	11.76	7.46	11.06	11.57	10.79	10.04	10.28	10.27	10.19
37	10.56	9.35	8.74	10.22	9.57	11.48	11.41	7.13	10.57	11.11	10.23	9.31	9.57	9.42	9.50
38	11.39	8.75	10.31	11.87	10.41	13.06	12.33	7.98	11.61	11.64	11.07	10.73	10.94	10.52	10.27
39	11.41	9.35	9.92	11.08	10.14	12.25	11.91	7.41	11.01	11.56	10.91	10.06	10.38	9.82	9.59
40	12.30	10.61	10.51	12.06	11.67	13.26	13.07	9.13	12.53	12.84	12.42	11.27	11.46	11.08	11.00
41	13.91	11.63	12.37	14.32	13.01	15.11	14.52	10.69	14.39	14.23	13.96	13.20	13.55	13.58	13.39
42	14.75	12.59	11.92	13.01	11.89	13.75	13.31	10.17	12.18	14.06	13.19	12.13	11.93	10.62	9.71
43	12.79	10.69	9.74	10.84	10.23	11.62	11.50	8.42	10.16	12.10	11.15	9.98	9.67	7.97	8.15
44	9.74	8.41	6.53	7.58	8.23	8.70	9.23	6.96	9.66	9.35	8.50	6.78	7.02	8.05	9.29
45	13.77	11.50	11.17	12.59	12.16	13.62	13.71	10.30	13.34	13.52	12.94	11.94	12.12	12.64	12.57
46	13.18	11.23	10.54	11.77	10.24	12.38	11.87	9.07	11.33	12.63	11.68	10.44	10.51	9.21	8.59
47	14.16	13.15	14.04	15.20	13.62	15.76	15.11	12.43	14.31	15.50	14.59	13.77	13.91	12.39	11.90
48	15.22	13.19	13.42	14.30	14.54	15.22	15.49	12.25	14.16	15.24	14.23	13.57	13.60	12.38	12.64

ANBAR, MICHAEL
353-30-1453

FIGURE 14

MEANS

GROUP =	L	C
VARIABLE		
1 M74	.795895	1.61160
2 M79	1.41739	3.29569
3 M79	.46739	1.64648
4 M80	.48348	1.02760
5 M85	1.33549	1.78188
6 M93	.53025	1.67206
7 M94	5.27349	31.50446
8 M95	2.97522	12.10458
9 M95	1.14343	3.22528
10 M102	1.97261	.78436
11 M103	2.45304	1.13528
12 M109	1.14957	3.53656
13 M112	6.35100	2.52328
14 M116	1.57435	.78566
15 M117	2.93423	1.35248
16 M118	1.14609	.70568
17 M121	.73153	1.51088
18 M136	1.73595	2.63588
19 M140	1.39391	1.92446
20 M142	1.02304	1.54559
21 M145	2.73217	5.35848
22 M146	1.12748	2.42668
23 M148	2.44522	4.15968
24 M150	1.70478	3.52598
25 M159	.50509	1.20648
26 M164	1.47087	2.79006
27 M168	14.78174	35.12006
28 M169	4.57795	9.41168
29 M175	7.88304	25.64286
30 M182	8.76174	22.64568
31 M183	1.51913	4.47528
32 M192	40.36133	5.72168
33 M193	21.20130	4.57526
34 M196	4.92304	28.85126
35 M197	1.78591	7.37568
36 M200	1.00133	1.51286
37 M203	.56957	1.21168
38 M204	1.53509	3.75766
39 M205	1.01652	1.92808
40 M207	1.35915	2.09006
41 M208	4.24739	16.15106
42 M209	1.49478	4.93768
43 M210	2.57479	4.90486
44 M218	2.28217	4.92646
45 M224	1.76911	3.92848
46 M225	1.18915	2.11206
47 M228	1.18739	1.94206
48 M231	.54391	1.02766
49 M250	1.31652	2.42166
50 M318	1.35067	.35168
51 M319	.34685	.33766
52 SUM	1.00000	2.00000
ACCOUNTS	23.	25.

FIGURE 15

STEP NUMBER 1
VARIABLE ENTERED 35 M197

VARIABLE	F TO REMOVE	FORCE LEVEL
	DF= 1 46	
35 M197	77.320	1

VARIABLE	F TO ENTER	FORCE LEVEL	TOLERANCE
	DF= 1 45		
1 M74	19.711	1	.828849
2 M78	23.689	1	.834129
3 M79	37.924	1	.782451
4 M80	23.301	1	.838190
5 M86	12.259	1	.870833
6 M93	13.014	1	.844535

STEP NUMBER 2
VARIABLE ENTERED 8 M35

VARIABLE	F TO REMOVE	FORCE LEVEL
	DF= 1 45	
8 M35	60.325	1
35 M197	97.112	1

VARIABLE	F TO ENTER	FORCE LEVEL	TOLERANCE
	DF= 1 44		
1 M74	.092	1	.564376
2 M78	.485	1	.615665
3 M79	3.932	1	.430895
4 M80	.591	1	.435774
5 M86	.323	1	.665307
6 M93	1.089	1	.879073
7 M94	6.318	1	.359350
9 M96	3.581	1	.373794
10 M102	7.859	1	.859730
11 M103	10.946	1	.864764
12 M109	.398	1	.884129
13 M112	7.464	1	.864145
14 M116	11.513	1	.843265

STEP NUMBER 15
VARIABLE ENTERED 47 M223

VARIABLE	F TO REMOVE	FORCE LEVEL
	DF= 1 32	
6 M33	4.245	1
8 M35	5.450	1
11 M103	13.027	1
14 M115	26.043	1
15 M117	1.353	1
23 M148	24.362	1
24 M150	3.379	1
28 M169	7.815	1
30 M182	5.727	1
31 M183	15.491	1
32 M192	18.823	1
33 M195	14.922	1
35 M197	33.259	1
47 M223	1.568	1
50 M318	20.179	1

VARIABLE	F TO ENTER	FORCE LEVEL	TOLERANCE
	DF= 1 31		
1 M74	.588	1	.323941
2 M78	.341	1	.293690
3 M79	.316	1	.130986
4 M80	.955	1	.224234
5 M86	.096	1	.487871
7 M94	.387	1	.208220
9 M96	.079	1	.191766
10 M102	.275	1	.188740
12 M109	.030	1	.274432
13 M112	2.975	1	.527261
16 M118	.972	1	.558126
17 M121	.142	1	.393252
18 M126	.232	1	.329842
19 M140	.374	1	.556038
20 M142	.485	1	.504500
21 M145	.030	1	.352794
22 M146	.201	1	.428966
25 M160	.353	1	.278788
26 M164	.051	1	.582123
27 M168	.230	1	.375410
29 M178	.471	1	.459193
34 M196	.253	1	.347733
36 M200	5.433	1	.394858
37 M203	2.837	1	.358768
38 M204	.011	1	.354150
39 M205	.069	1	.527018
40 M207	.073	1	.471641
41 M210	.591	1	.435093
42 M214	.102	1	.247341
43 M215	.342	1	.441451
44 M216	.[illegible]	1	.[illegible]
45 M220	.[illegible]	1	.[illegible]
46 M222	.[illegible]	1	.[illegible]
48 M231	2.479	1	.[illegible]
49 M[illegible]	2.[illegible]	1	.[illegible]
51 M[illegible]	1.[illegible]	1	.[illegible]

WILKS' LAMBDA .[illegible] DEGREES OF FREEDOM 1 1 46
APPROXIMATE F-STATISTIC [illegible] DEGREES OF FREEDOM 15.00 32.00

FIGURE 16

STEP NUMBER 49
VARIABLE ENTERED 45 M224

VARIABLE	F TO REMOVE	FORCE LEVEL		VARIABLE	F TO ENTER	FORCE LEVEL	TOLERANCE
	DF= 1 17				DF= 1 16		
2 M78	3.954	1		1 M74	.236	1	.193106
3 M79	1.729	1		4 M83	.025	1	.134387
7 M94	32.721	1		5 M86	.002	1	.156815
10 M112	6.577	1		6 M93	.066	1	.042416
11 M113	50.118	1		8 M95	.001	1	.120866
15 M117	12.361	1		9 M96	.293	1	.031415
16 M118	3.562	1		12 M109	.015	1	.130397
18 M135	3.941	1		13 M112	.013	1	.257906
19 M140	25.004	1		14 M116	.046	1	.088961
20 M142	9.267	1		17 M121	.131	1	.120778
23 M149	12.870	1		21 M145	.100	1	.085782
24 M150	5.801	1		22 M146	.362	1	.117529
25 M155	7.020	1		26 M164	.204	1	.287412
28 M169	136.930	1		27 M168	.204	1	.064119
29 M175	2.053	1		30 M182	.003	1	.107116
31 M189	34.225	1		38 M204	.213	1	.089256
32 M192	11.075	1		39 M205	.143	1	.068208
33 M195	16.380	1		42 M209	.139	1	.017539
34 M196	6.365	1		44 M218	.004	1	.053036
36 M197	103.500	1		46 M225	.108	1	.101219
37 M200	46.529	1		51 M319	.247	1	.160053
38 M204	30.551	1					
40 M207	56.831	1					
41 M208	33.473	1					
42 M210	17.779	1					
44 M224	1.330	1					
47 M226	56.929	1					
48 M231	14.578	1					
49 M250	16.769	1					
50 M310	74.090	1					

SU-STATISTIC OF WILKS' LAMBDA .0023294 DEGREES OF FREEDOM 50 1 46
APPROXIMATE F-STATISTIC 242.701 DEGREES OF FREEDOM 30.00 17.00

Figure 17 presents the distances from each group centroid to the points representing each of the cases calculated from the 30 variables finally selected. This table gives also the probability of assignment of each case to a given group. One can see that each case in the L group have been assigned to it with a probability of unity and similarly each case in the C group (controls) has been assigned to this group with a probability of unity. Figure 18 is a graphical presentation of the same data recalculated for 2 dimensional projection. The digits 1 and 2 are the positions of the centroids of each group respectively. The resolution of the computerized printout is limited so that if more than one case fall on the same overall unit area on the plot they will be presented by just a single mark. Therefore, only 19 points are shown for the pathological sample and coincidentally 19 points were printed for the controls. The actual coordinates for each of the 48 cases is presented in Figure 17.

We see that the stepwise discriminant analysis has separated the cases in a more "decisive" manner than by the WNI test (Figure 11). This is not surprising in view of the fact that it was given the "best" preselected 50 variables and its own procedure used only 30 out of these for the ultimate classification. Although the quantitative and graphical presentations by this procedure are the most "convincing" among the 3 types of classification, it is the clustering analysis that demonstrated in an utterly unbiased fashion that we had in this study just <u>two</u> biochemically distinct groups. As stated elsewhere in this report we shall proceed with additional comparative statistical analyses of other series of biological samples, including sets less distinctly different than the one shown here, before deciding which statistical treatment is most suitable for a given problem.

FIGURE 17

OGROUP	CASE	L		C	
L	1	31.5	1.000	1665.3	0.000
	2	32.5	1.000	1629.6	0.000
	3	19.5	1.000	1518.7	0.000
	4	24.1	1.000	1910.9	0.000
	5	25.7	1.000	1654.6	0.000
	6	31.2	1.000	1484.1	0.000
	7	27.5	1.000	1711.8	0.000
	8	27.5	1.000	1683.4	0.000
	9	25.8	1.000	1667.8	0.000
	10	29.6	1.000	1710.2	0.000
	11	22.2	1.000	1745.5	0.000
	12	15.8	1.000	1536.8	0.000
	13	17.1	1.000	1696.4	0.000
	14	21.6	1.000	1721.7	0.000
	15	25.4	1.000	1642.3	0.000
	16	21.0	1.000	1605.8	0.000
	17	19.4	1.000	1704.1	0.000
	18	29.8	1.000	1616.1	0.000
	19	23.9	1.000	1633.8	0.000
	20	15.5	1.000	1737.9	0.000
	21	29.5	1.000	1705.8	0.000
	22	25.7	1.000	1673.6	0.000
	23	23.0	1.000	1696.5	0.000
OGROUP	C	L		C	
C	24	1727.1	0.000	38.4	1.000
	25	1647.0	0.000	32.6	1.000
	26	1825.0	0.000	30.2	1.000
	27	1599.3	0.000	27.9	1.000
	28	1755.4	0.000	35.4	1.000
	29	1607.8	0.000	31.1	1.000
	30	1675.4	0.000	28.5	1.000
	31	1621.8	0.000	34.3	1.000
	32	1629.7	0.000	25.6	1.000
	33	1735.8	0.000	35.3	1.000
	34	1629.4	0.000	28.7	1.000
	35	1699.2	0.000	28.3	1.000
	36	1599.2	0.000	28.5	1.000
	37	1612.0	0.000	26.8	1.000
	38	1752.4	0.000	26.3	1.000
	39	1561.0	0.000	31.8	1.000
	40	1651.5	0.000	31.3	1.000
	41	1732.5	0.000	36.3	1.000
	42	1632.9	0.000	23.5	1.000
	43	1751.8	0.000	16.8	1.000
	44	1612.0	0.000	33.4	1.000
	45	1727.2	0.000	33.5	1.000
	46	1675.2	0.000	35.4	1.000
1					
	47	1635.8	0.000	37.5	1.000
	48	1599.8	0.000	42.6	1.000

POINTS TO BE PLOTTED

OVARIABLE	OGROUP	MEAN COORDINATES		SYMBOL FOR CASES	SYMBOL FOR MEAN
2 V78					
3 V79	L	21.12	.00	L	1
7 V84	C	-19.43	.00	C	2
10 V102					
11 V103					
15 V117					
16 V118					
18 V136					
19 V140					
20 V142					
23 V148					
24 V150					
25 V159					
28 V169					
29 V175					
31 V183					
32 V192					
33 V193					
34 V196					
35 V197					
36 V200					
37 V203					
40 V207					
41 V208					
42 V210					
45 V224					
47 V228					
48 V231					
49 V250					
50 V318					

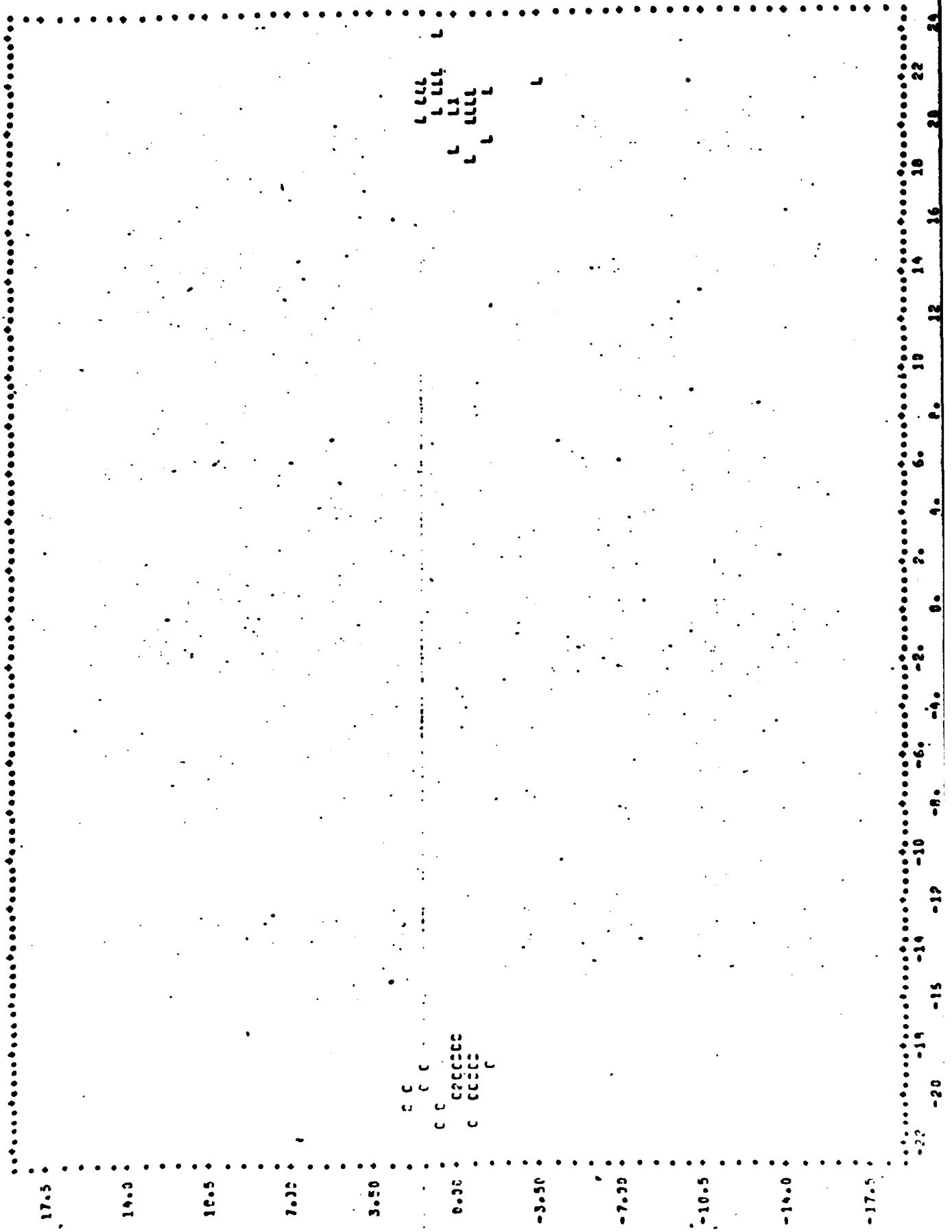
OGROUP L

CASE	X	Y	CASE	X	Y	CASE	X	Y
1	20.99	.85	11	22.09	.69	21	21.51	1.72
2	20.56	-.80	12	19.50	-1.10	22	21.16	1.06
3	19.40	.19	13	21.57	-.41	23	21.42	-1.22
4	24.11	.38	14	21.80	-3.37			
5	20.83	.19	15	20.80	-.59			
6	18.76	-.46	16	20.38	1.56			
7	21.49	-.40	17	22.36	.65			
8	21.14	1.31	18	20.46	-.57			
9	21.08	-.65	19	20.69	-.71			
10	21.57	.83	20	22.04	1.06			

OGROUP C

CASE	X	Y	CASE	X	Y	CASE	X	Y
24	-19.73	-.87	34	-19.57	-.40	44	-18.62	-1.22
25	-18.32	-.19	35	-17.54	-.48	45	-20.04	.33
26	-21.28	.75	36	-19.52	-.06	46	-19.38	-.61
27	-17.78	-.35	37	-21.18	-1.04	47	-19.49	1.09
28	-20.44	1.50	38	-20.19	-.13	48	-18.95	-.50
29	-18.50	-.72	39	-15.01	-.26			
30	-19.47	2.04	40	-19.17	-.32			
31	-18.85	-.17	41	-21.37	.12			
32	-18.87	.15	42	-19.00	1.16			
33	-20.11	-.78	43	-21.52	.25			

FIGURE 18


17.5
14.0
10.5
7.00
3.50
0.00
-3.50
-7.00
-10.5
-14.0
-17.5
-22
-20
-18
-15
-14
-12
-10
-8.
-6.
-4.
-2.
0.
2.
4.
6.
8.
10
12
14
16
18
20
22

4. EXPERIMENTAL RESULTS

A. Hospital Patient Studies

The entire multicomponent mixture analysis procedure has been tested using four sets of pathological and three sets of control samples. Two sets of samples representing liver malfunction have been obtained. One set of eight samples includes a variety of diseases, including malignancies, with liver involvement as a primary or secondary reason for hospitalization. Another set of 12 samples is more homogeneous representing primarily alcoholic hepatitis. A sample set from 7 individuals hospitalized in the same ward as the alcoholic hepatitis patients was evaluated as a control set.

Samples were provided by the staff of Childrens Hospital from 7 pneumonia cases, 12 diarrhea patients diagnosed as of viral origin, and 6 samples selected from patients hospitalized for concussion and tonsilectomy, situations unlikely to involve an infectious organism. An additional set of 13 adult control samples was obtained from healthy volunteers within the university.

Each sample was analyzed in duplicate and control and pathological samples were interspersed. Summed spectra were assembled into data sets for processing in the CYBER by either the Wilcoxon WNI programs or the BMDP programs.

The results of the Wilcoxon-WNI program are presented for several data sets in figures 19 through 25. The graphs display WNI(1) versus WNI(2) for each sample in both groups. The WNI(1) axis represents the difference of a given sample from the average spectra of the pathological group while the WNI(2) axis shows the difference of a spectra from the control group class average. In this representation control samples should fall close to the vertical axis and a maximum distance from the horizontal, while pathological samples should fall near the horizontal axis and away from the vertical axis.

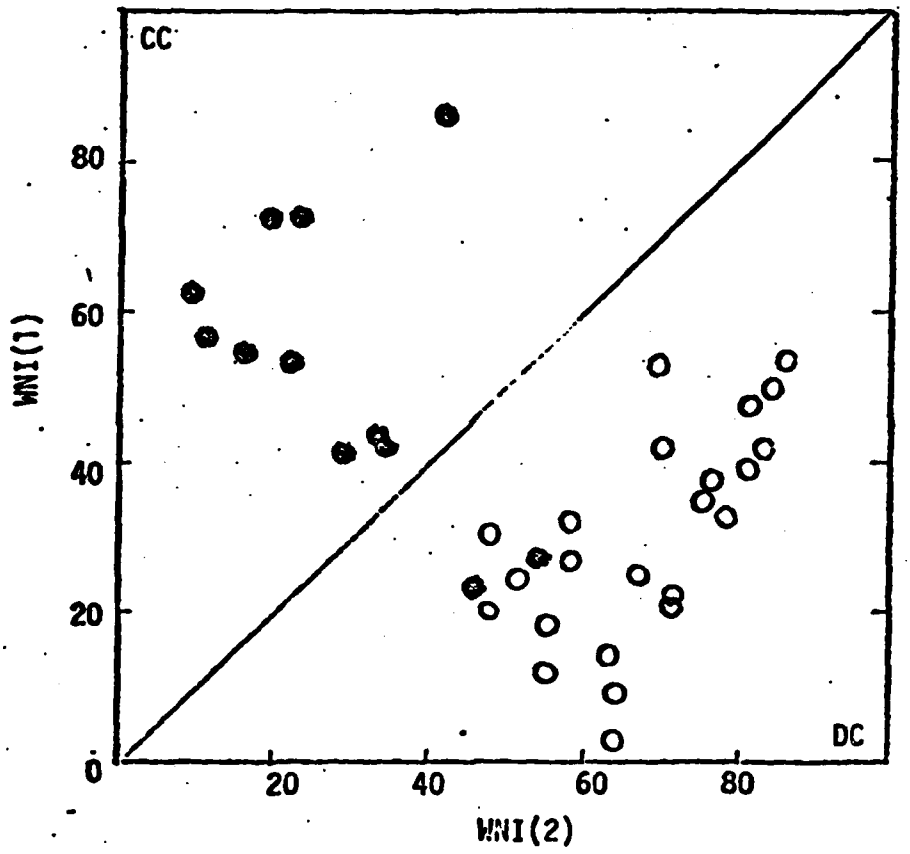


FIG.1g WNI PLOT OF CHILDREN'S CONTROL SAMPLES (CC-CLOSED CIRCLES) VS DIARRHEA IN CHILDREN (DC-OPEN CIRCLES) USING ALL m/e VALUES WEIGHTED BY 1/p FOR COMPUTING WNI'S

ANBAR, MICHAEL
353-30-1453

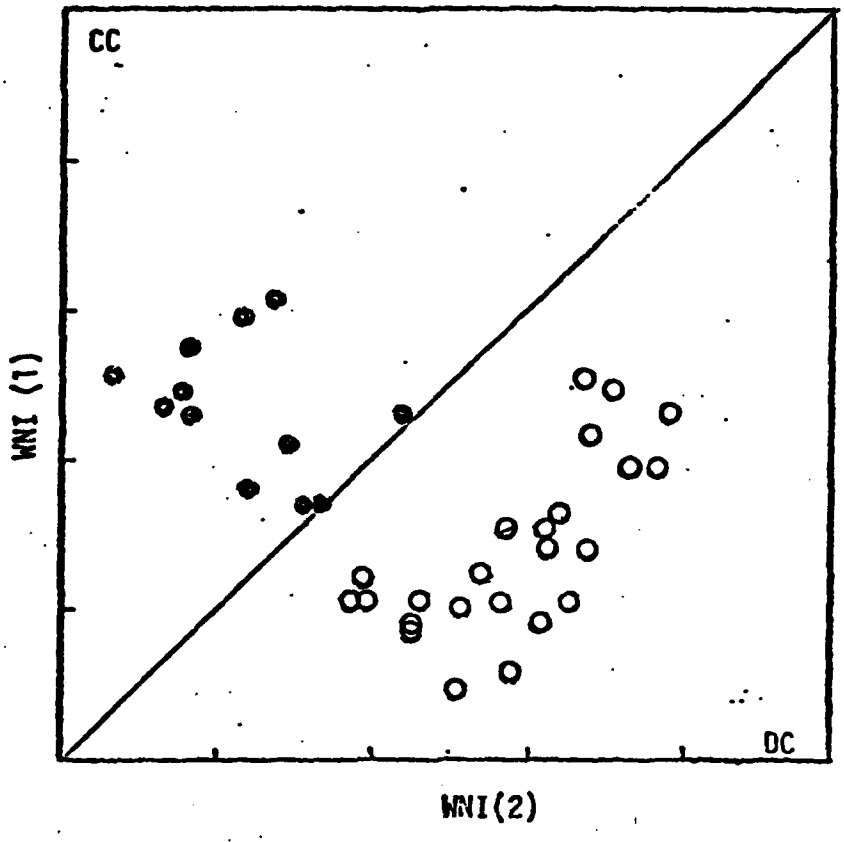


FIG.20 CONTROL (CC-CLOSED CIRCLES) VS DIARRHEA SAMPLES (DC-OPEN CIRCLES) USING 8 m/e VALUES OF LOWEST p VALUES TO COMPUTE WNI'S

ANBAR, MICHAEL
353-30-1453

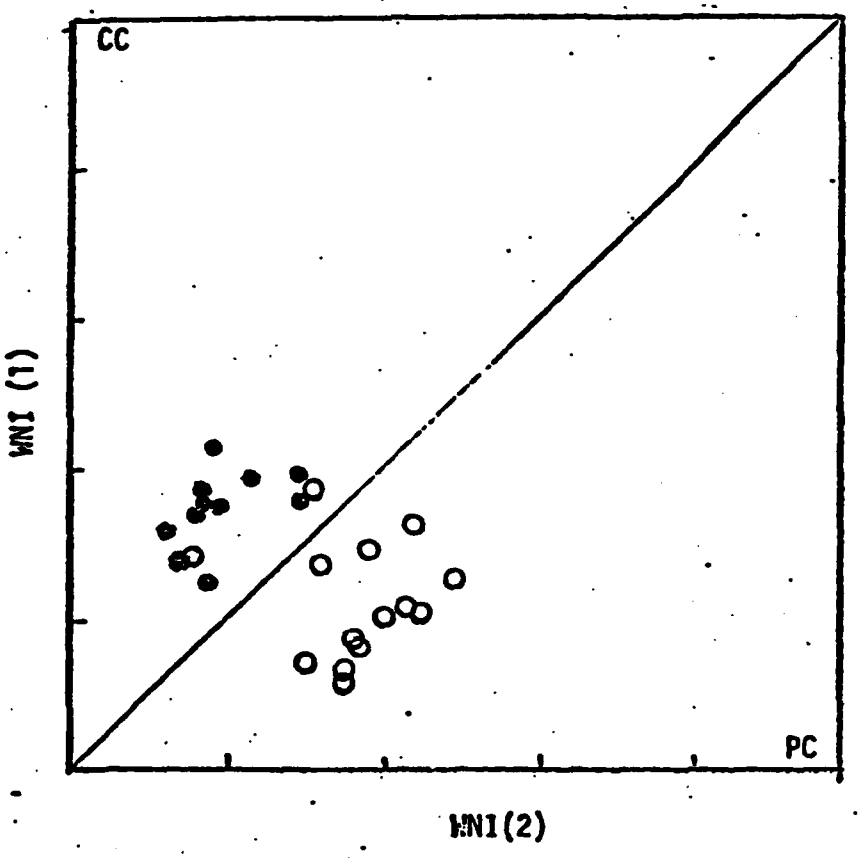


FIG. 21 CHILDREN'S CONTROL SAMPLES (CC-CLOSED CIRCLES) VS CHILDREN'S PNEUMONIA SAMPLES USING 14 m/e VALUES FOR WNI'S (PC-OPEN CIRCLES).

ANBAR, MICHAEL
353-30-1453

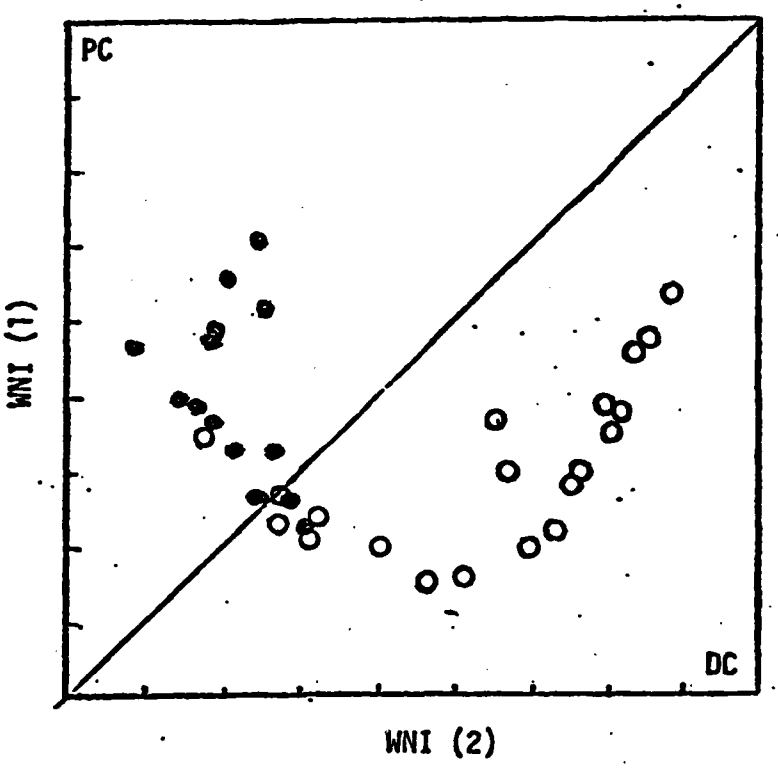


FIG. 22 PNEUMONIA SAMPLES (PC-CLOSED CIRCLES) VS DIARRHEA SAMPLES (DC-OPEN CIRCLES)

ANBAR, MICHAEL
353-30-1453

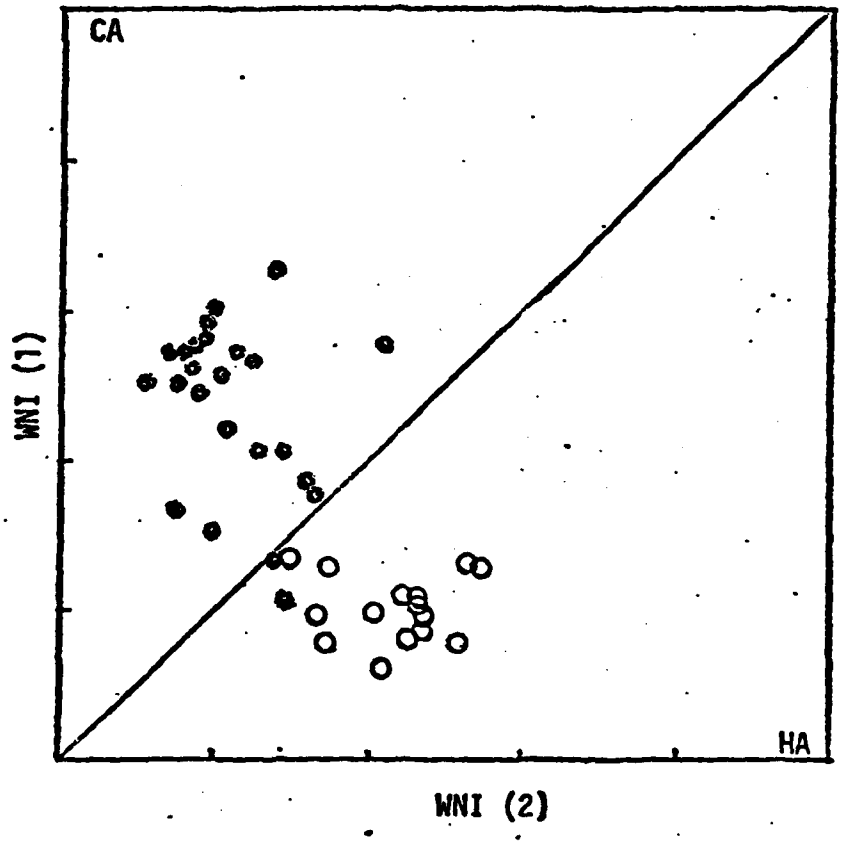


FIG. 23 ADULT CONTROL SAMPLES (CA CLOSED CIRCLES) VS MIXED LIVER DISORDERS (HA-OPEN CIRCLES) USING ALL m/e VALUES.

ANBAR, MICHAEL
353-30-1453

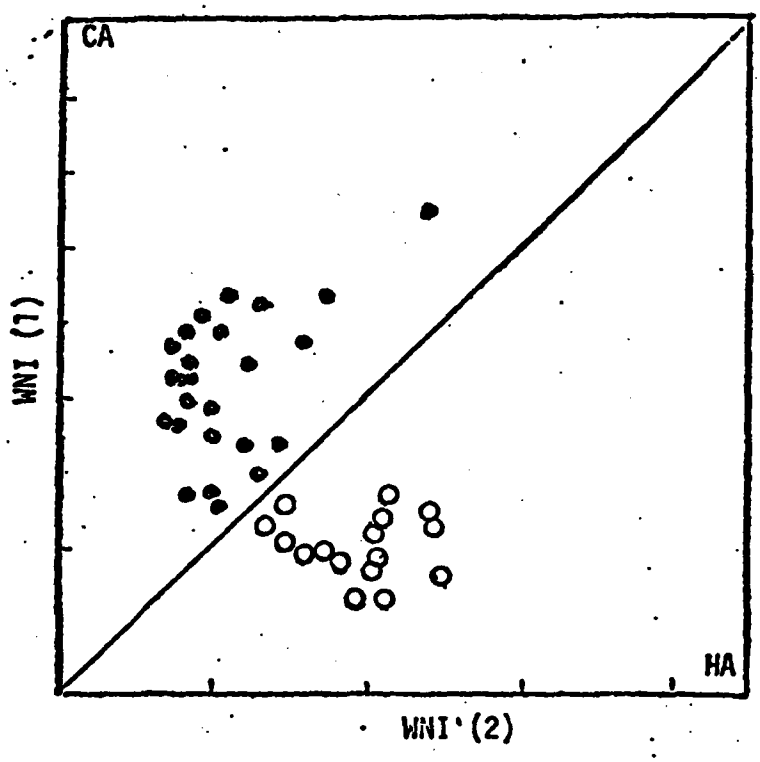


FIG. 24 ADULT CONTROL SAMPLES (CA-CLOSED CIRCLES) VS MIXED LIVER DISORDERS (HA-OPEN CIRCLES) USING 12 m/e VALUES.

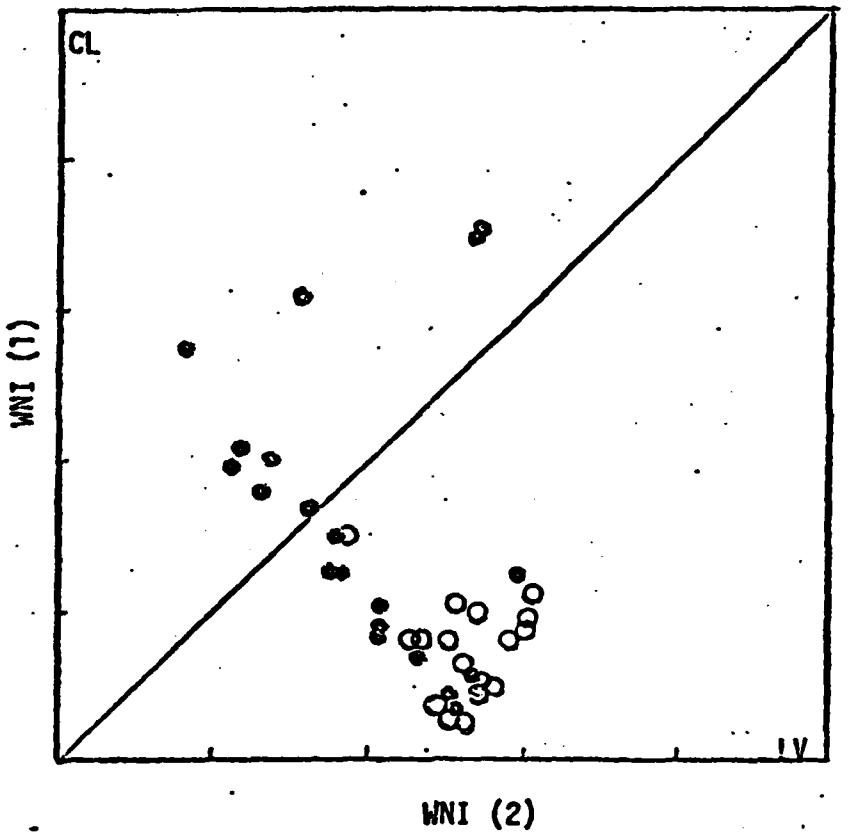


FIG. 25 HOSPITALIZED CONTROLS (CL-CLOSED CIRCLES) VS ALCOHOLIC LIVER DISEASE (LV-OPEN CIRCLES)

The comparison of the diarrhea in children with controls (Fig. 19) shows two control samples falling within the pathological group when all masses are used with a weighting function of 1/p to compute the WNI's. The use of only the eight mass of lowest p-value (Fig. 20) not only shows complete separation of control and diarrhea cases, but seems to reveal a distinct grouping among the diarrhea cases that may be the result of a different infecting organism. For several of the diarrhea samples collected within a few days of each other rhodovivus was detected by independent tests.

The pneumonia cases show one individual from the pathological set falling within the control group using WNIs based on the 14 peaks of lowest p-value (Fig. 21). The hospital records for this individual indicated that this patient was discharged the day after the sample was taken, so it might be surmised that this patient had already recovered from the infection. Adding this individual to the control group and removing the sample pair from the pathological group did not appreciably change the clustering of samples obtained before, indicating that this sample properly belongs in the control group.

An example of differential diagnosis is given in Fig. 22 comparing the diarrhea to the pneumonia cases. Incomplete separation was obtained using all masses weighted as 1/p in the WNI calculation.

Figures 23 and 24 show the mixed set of liver disorders compared to healthy adult controls, using all masses weighted by 1/p and using only the 12 masses of lowest p-value. Again the use of only the diagnostic peaks improves the separation between the two classes. Comparing the alcoholic hepatitis set with hospital controls (Fig. 25) reveals a weakness in the choice of this control set, so that although the alcoholic hepatitis samples are tightly clustered, a number of samples from the control set fall within this

grouping. In this experiment patients with gastric ulcer, functional gall bladder disease, and anemia appear to have similar profile patterns to those with liver disorders. This example illustrates the importance of obtaining well characterized samples upon which to base the initial pattern analysis. Comparing the alcoholic hepatitis set with the adult controls (Fig. 11) shows a much clearer separation. In both comparisons the set of pathological samples appears to be a well defined homogenous group. The use of the adult control set in the comparison generates a different set of masses with the lowest p-values. In addition the p-values were 2 to 3 orders of magnitude smaller using the adult controls. Tabular results of the Wilcoxon-WNI analysis of these samples is presented in Figure 10.

These two groups of samples were also analyzed using the BMDP programs. First the BMDP 3D program was used to calculate the t-statistic, separate and pooled, for the peak intensities at each m/e value for the null hypothesis of equivalent means for both the pathological and control groups. Based on these results 51 masses were selected. A significant number of the masses selected by this method also show low p-values by the Wilcoxon test. Using these selected variables both cluster and stepwise discriminant analysis programs were used to classify the individual cases. These results have been described in section 3D.

B. Longitudinal Studies on Virus Infected Patients.

Controlled longitudinal studies were carried out on two sets of volunteers at the AFRIID. One group consisted of seven individuals who received live virus vaccine for sandfly fever. Two additional individuals in this group received a placebo injection. None of the participants were told whether they received the vaccine or control injection. Morning urine samples were collected from participants 4 days prior to the injection, the day of the

injection, for 8 consecutive days subsequent to injection and finally 28 days after the injection.

A second volunteer group was innoculated with Dengue fever live virus vaccine provided morning urine samples 3 days prior to and for 21 days following injection. All samples were stored frozen without preservative, shipped to this laboratory packed in dry ice, and kept frozen until analysis. We initially analyzed samples from three individuals in the sandfly fever experiment and two participants of the Dengue fever group without any prior knowledge of the sample classification (vaccine vs. control) or the expected time and duration of the symptoms. The samples from each individual were prepared and analyzed in duplicate in a random sequence. In a few cases during this series a sample had to be repeated due to instrument malfunction during data acquisition. In these cases we found that better replicate samples were obtined by using the salt saturated urine remaining from the previous extraction rather than using the original refrozen sample. We suspect that a bacterial contamination may have affected the original samples during their exposure to room temperature. We do not observe similar differences in repeated analyses of samples that are collected and stored with $ZnSO_4$ as a preservative.

The data from the first three individuals of the sandfly fever showed an extreme difference in the patterns of two of the individuals (Joffe, LeBlanc) compared to the third (Berry).

Throughout the sampling period Berry's profiles show significantly lower intensity over the entire mass range. This individual also apparently did not consume any caffeinated beverages during the study, leading to a further qualitative difference in this person's pattern compared to the other two participants. We, therefore, did not include Berry's patterns in the initial

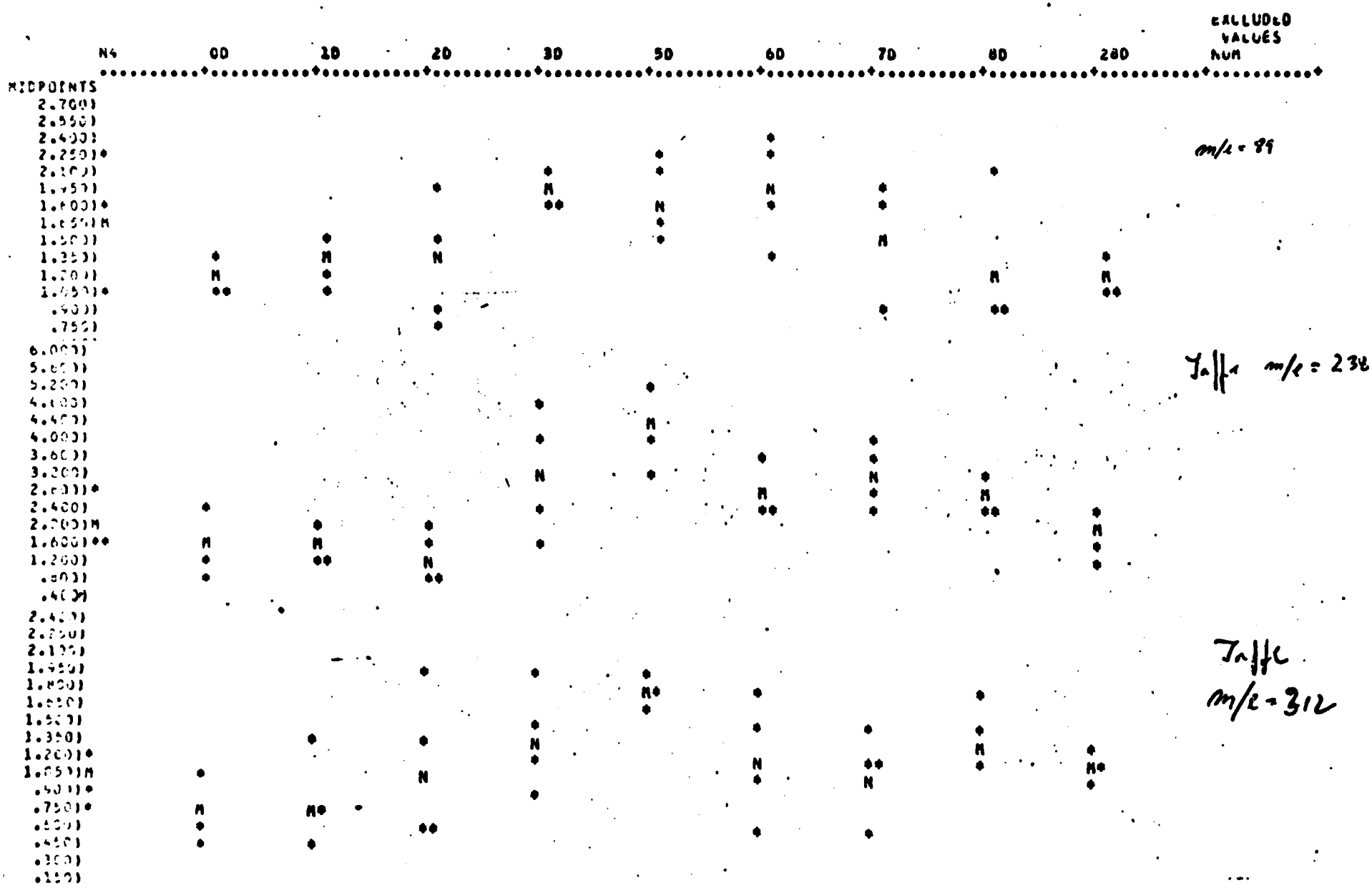
analysis of the data. (We have received the information that this individual served as one of the controls after the analysis of all samples were completed). With only two individuals to examine we were concerned that dietary variations plus individual variations in the response to the vaccine might obscure the detection of the pathological pattern or location of the maximum response period. Two approaches were partially successful in handling this problem of deciphering an unknown pattern appearing at an unknown time in a small data set.

In the first case the BMDP 7D program for variable stratification was employed to show the means and intensity distributions for each m/e value as a function of the sample collection sequence. An example is shown in Figure 26. This analysis, although tedious and inefficient for routine use, did reveal a number of masses showing changes in intensity distributions at days 3, 4 and 5. A second approach involved comparing patterns of days minus 4 and zero against days 3, 4, and 5.

Selection of days 3-5 was partly based on the previous stratification of variable study and also on the expected time for development of symptoms for the particular virus employed. Again BMDP 3D was used for a t-test on each m/e value between the selected groups of days. Using the variables selected in this manner the BMDP 7M discriminant analysis program was used to derive the classification function for the same two groups (days -4, 0 vs. 3, 4, 5). In addition this function was used to classify days 2 as a third group and days 7, 8 and 28 as a fourth group.

Figure 27 shows the results of this classification. The most obvious result to be noted is that the "normal" group (day -4 and 0) separates completely from the presumed pathological group (days 3-5). Of more interest is the classification of the group based on day 2 showing a pathological

FIGURE 26



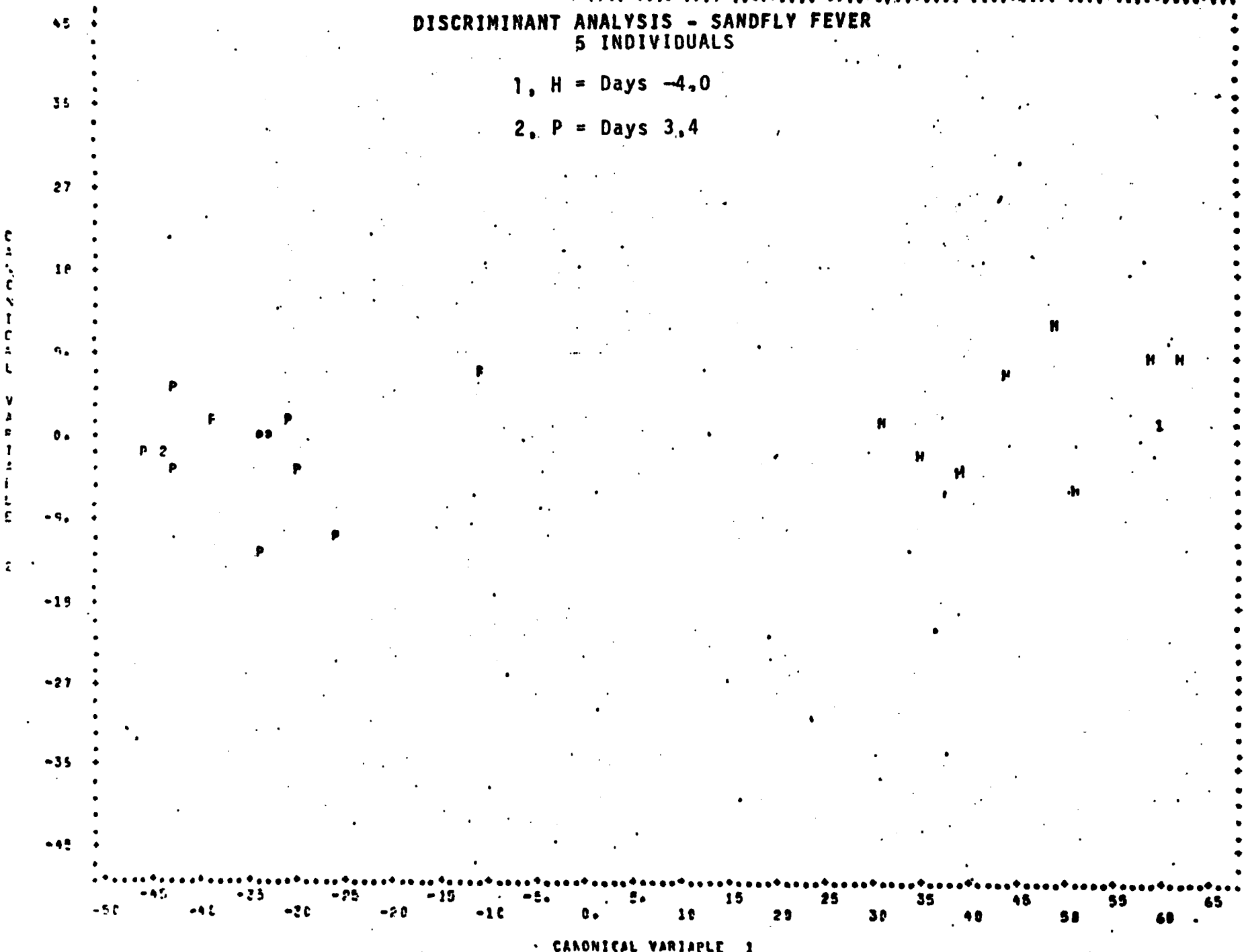
pattern for Joffe and a normal pattern for LaBlanc. In the final group, day 28 shows a return to the normal pattern for both individuals with days 7 and 8 showing predominantly a pathological pattern.

We have been informed by AFRIID that neither subject developed any clinical symptoms before the middle of Day 3, thus since these were morning urines we have three urine samples (2 of Joffe and 1 of LeBlanc) that showed the pathological pattern prior to manifestation of clinical symptoms. Next we are told that by day 7 both subjects were feeling well, free of symptoms. But still on day 8 LeBlanc shows an unambiguous pathological pattern, whereas Joffe shows an ambiguous result. Unfortunately in this series no samples were collected beyond day 8 so that it is impossible to determine at this point the time of unambiguous disappearance of the pathological pattern. In any case it seems to appear prior to the clinical symptoms and persist beyond the time of their disappearance.

The samples from the remaining participants in this study were recently shipped to us and days -4, 0, 3, 4 and 5 from persons receiving the vaccine were selected for initial analysis. During the preparation of this set the addition of EDTA was inadvertently omitted from the extraction procedure. Our statistical analysis of this sample set shows a detectable pattern difference due to this change in the procedure making a direct classification of all seven vaccinated individuals more complicated. The use of the t-test comparing normal and "pathological" periods of this last set of samples leads to a selection of variables that correctly classifies these samples using the discriminant analysis program (see Figure 28). The use of these variables with the discriminant analysis program to classify samples from all seven individuals into four groups was also successful; still we plan to repeat the analysis of the first two individuals without using EDTA in the extraction procedure to eliminate this this artifact from the classification.

FIGURE 27-
DISCRIMINANT ANALYSIS SANDFLY FEVER

	INCORRECT CLASSIFICATIONS	MAHALANOBIS D-SQUARE FROM AND POSTERIOR PROBABILITY FOR GROUP =				
GROUP A		A		C		
CASE						
1		14.4	1.000	******	0.000	
2		14.4	1.000	******	0.000	
3		14.4	1.000	******	0.000	
4		14.4	1.000	******	0.000	
5		14.4	1.000	******	0.000	
6		14.4	1.000	******	0.000	DAYS -4, 0; Joffe and LeBlanc
7		14.4	1.000	******	0.000	
8		14.4	1.000	******	0.000	
GROUP B		A		C		
CASE						
9	C	******	0.000	******	1.000	Cases 9,10 Day 2 Joffe
10	C	******	0.000	******	1.000	
11	A	******	1.000	******	0.000	
12	A	******	1.000	******	0.000	Cases 11; 12, Day 2 LeBlanc
GROUP C		A		C		
CASE						
13		******	1.000	14.4	1.000	
14		******	1.000	14.4	1.000	
15		******	1.000	14.4	1.000	
16		******	1.000	14.4	1.000	
17		******	1.000	14.4	1.000	
18		******	1.000	14.4	1.000	
19		******	1.000	14.4	1.000	Days 3, 4 , 5 Joffe, LeBlanc
20		******	1.000	14.4	1.000	
21		******	1.000	14.4	1.000	
22		******	1.000	14.4	1.000	
GROUP C		A		C		
CASE						
23	A	******	1.000	******	0.000	
24	A	******	1.000	******	0.000	Cases 23-25, Day 7,8,28 Joffe
25	A	******	1.000	******	0.000	
26	C	******	0.000	******	1.000	Cases 26-28, Days 7,8,28 Joffe
27	A	******	1.000	******	0.000	
28	A	******	1.000	******	0.000	Cases 29-31, Days 7,8,28 LeBlanc
29	C	******	0.000	******	1.000	
30	C	******	0.000	******	1.000	Cases 32-3 , Days 7,8,28 LeBlanc
31	A	******	1.000	******	0.000	
32	C	******	0.000	******	1.000	
33	C	******	0.000	******	1.000	
34	A	******	1.000	******	0.000	



BMDP7M - STEPWISE DISCRIMINANT ANALYSIS.
HEALTH SCIENCES COMPUTING FACILITY
UNIVERSITY OF CALIFORNIA, LOS ANGELES

FIGURE 28

PROGRAM REVISED JULY 7, 1975
MANUAL DATE -- 1975

The more interesting question is, however, whether the 5 additional subjects will exhibit a similar behavior to the first 2 subjects on day 2 and on days 7 and 8 and whether no effect will be found in the placebo cases. These results will be reported in next year's report.

C. Study of Human Tissue Culture Infected With Polio Virus.

This study was undertaken to determine the differences between infected and uninfected cell cultures and to see how soon after infection of the culture such differences can be detected. We have begun this study with a cell line of human embryonic lung and the Mahoney strain of polio virus. The tissue culture work was done by Dr. Howard Faden and his associates at the Children's Hospital in Buffalo.

The human embryonic lung culture was originally established by Dr. Fishaut of Denver by 7 to 8 passes of tissue obtained from an aborted fetus. These cells were grown in Eagles minimal essential growth medium which was supplemented with 5% newborn calf serum. Cultures with equal numbers of cells were prepared in 5 ml plastic culture tubes that were incubated at $37^{o}C$ for zero, 6, and 24 hours. For each of these three time points, nine tubes were prepared. Three tubes were left uninfected, three were infected with 1/10 ml of viral solution with a tissue culture infective dosage ($TCID_{50}$) of 1×10^{2}, and three tubes with a TCID of 1×10^{4}.

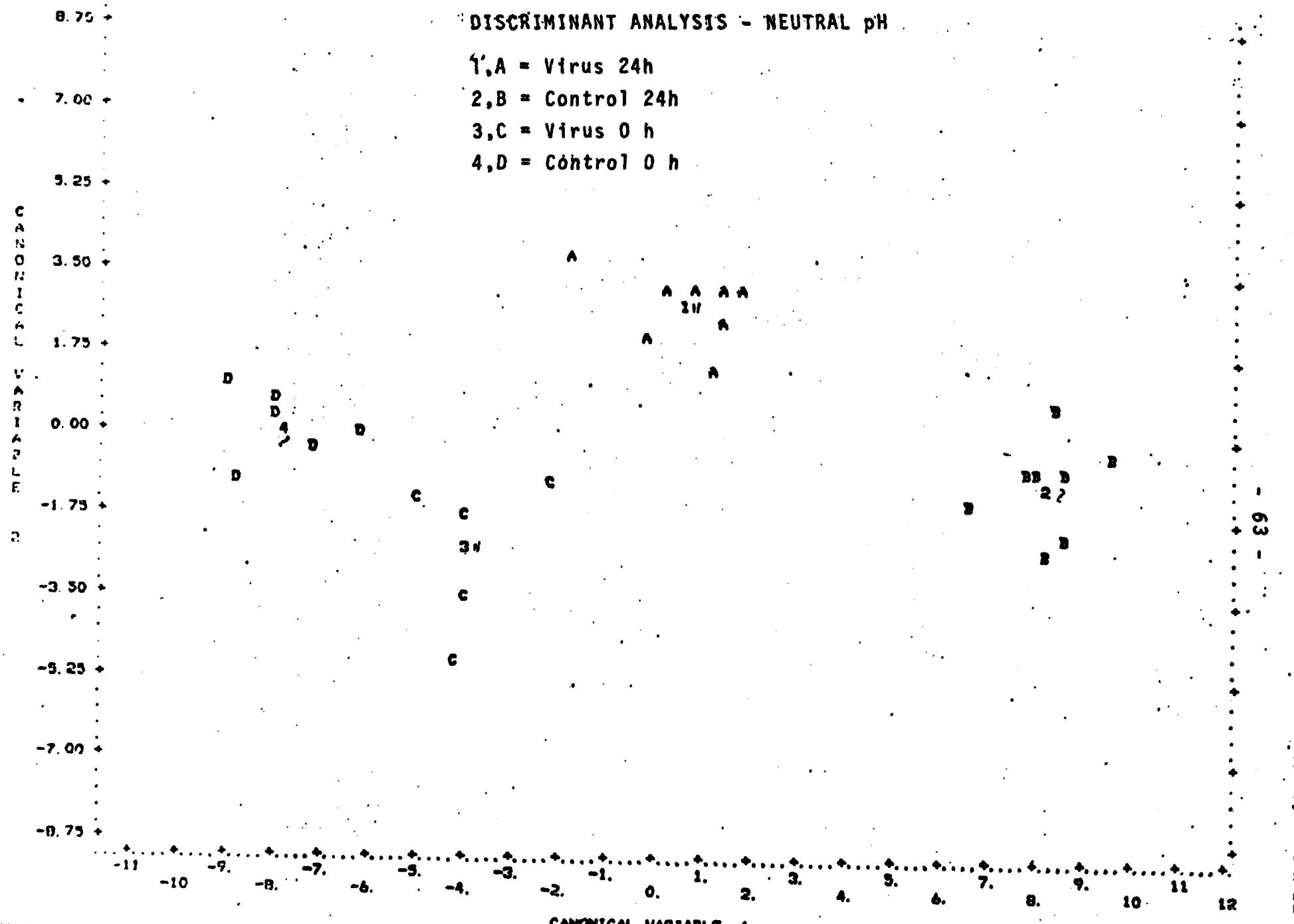
These titers of virus were allowed to adsorb onto the cells for one hour. At that time the cells were washed with isotonic phosphate buffer solution to remove excess virus particles. Following the wash, medium with newborn calf serum was added to the tubes, and the cells were incubated for the times indicated. When the incubation times were reached, the tubes were refrigerated and centrifuged to separate the cells. We received the chilled, cell free, supernate of this centrifugation.

To one ml of this medium we added .25 ml of cold 70% perchloric acid and .25 ml of cold 11 M KOH to precipitate proteins of the medium. To insure inactivation of the virus, 20 microliters of .3M $HgCl_2$ was added to each tube. The supernate of each tube was removed and diluted with an equal volume of distilled H_2O to facilitate pH adjustment. This diluted medium was then titrated with KOH and $HClO_4$ to a pH of 2, 7, or 10. A 150 microliter sample was removed at each pH and placed in a 1 cm by .3 cm glass culture tube which contained a folded 3 cm x 1 mm strip of Whatman fiberglass paper GP/C. The samples were evaporated onto this paper by blowing warm, dry nitrogen over them. The paper with the sample dried onto it, was then introduced into the mass spectrometrer.

Initially high and zero virus samples incubated for zero and 24 hours from each of the 3 tubes were prepared at pH 2, 7, and 10 in duplicate. The data were initially examined using the t-test to find peaks showing differences at 24 hours between zero and high virus. Using masses selected in this manner, cluster analysis and discriminant analysis) programs were used to classify the four groups consisting of 24 hours high and zero virus and 0 hours high and zero virus. A superior separation among these groups was observed for the samples prepared at pH 2 (see Figure 29 through 31). Based on this analysis pH 2 was selected for preparation of high and low virus samples from 6 hours and 24 hours.

The initial classification into four groups revealed several trends in the patterns. First there was a small but consistent difference in the zero time patterns of different virus concentrations indicating a change in the composition of the media due to the adding of virus. This is an artifact not normally considered in cell culture experiments and it can be eliminated in

FIGURE 30

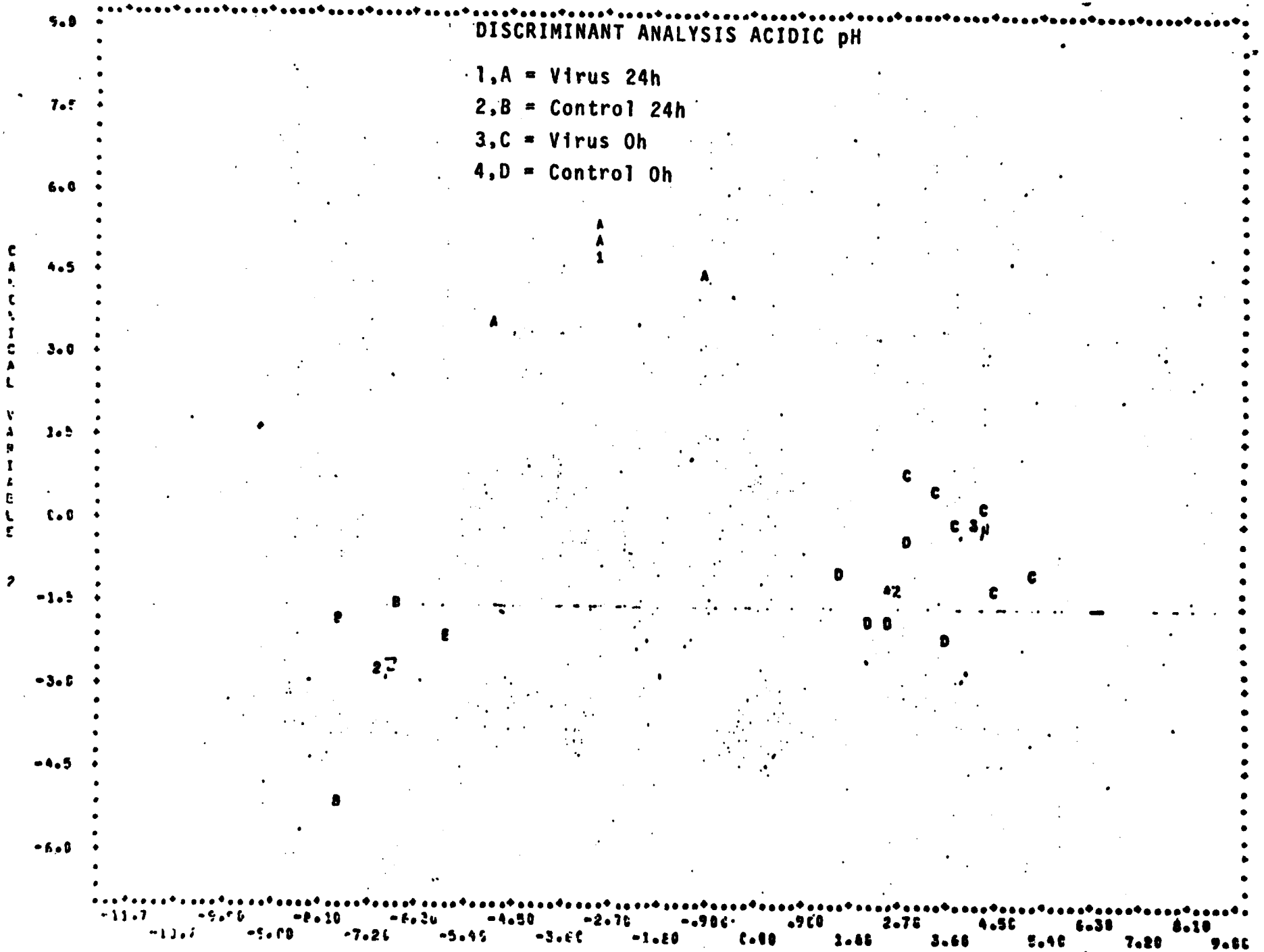


- 63 -

ANBAR, MICHAEL

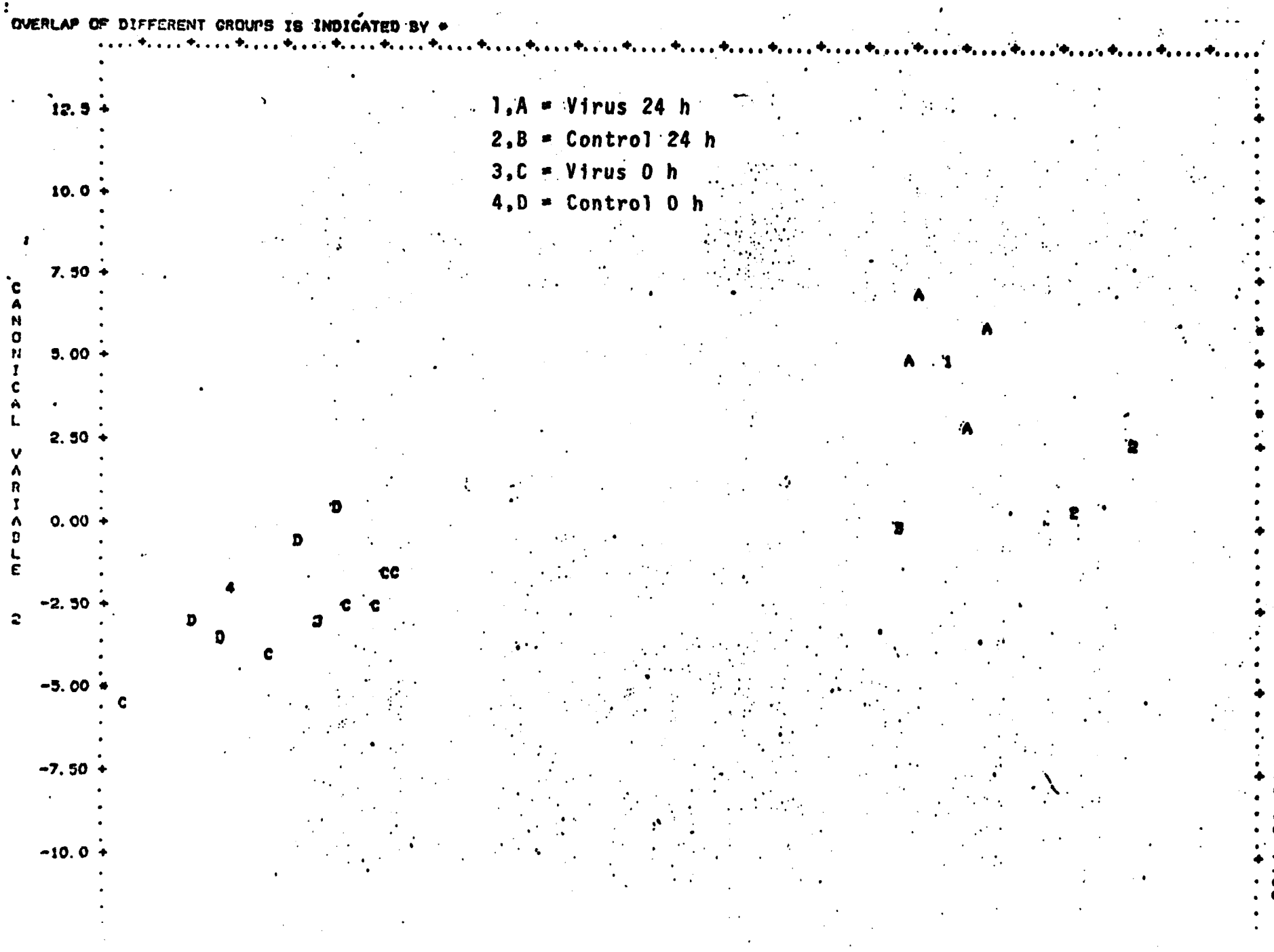
10MUPYM - STEPWISE DISCRIMINANT ANALYSIS.

FIGURE 29



ANBAR, MICHAEL

FIGURE 31
DISCRIMINANT ANALYSIS - BASIC pH



future studies by making all virus innoculations including the zero virus using a common media. Second, the incubation time and cell growth leads to time dependent changes in the intensity of selected m/e values. We observe both increases, which may be due to cell metabolites excreted into the media, and decreases possibly reflecting depletion of nutrients in the media. The third detectable trend is for the virus infected tubes to show similar time dependent trends but of reduced magnitude consistent with an inhibition in the effective cell growth rate. That is, the same peaks which increase or decrease in uninfected cultures show a smaller increase or decrease respectivly in the infected cultures. Finally and more interesting there are a limited number of m/e's showing intensity changes that may specifically reflect virus activity. These include cases where virus infected cultures show larger changes in the same direction observed in non-infected tubes.

Two additional t-tests were used to obtain new groups of m/e values. By performing the t-test on zero versus 24 hours, at high virus concentration, a set of peaks was obtained reflecting both time and virus dependent pattern changes.

Similarly a set of zero time media composition dependent peaks was found using the t-test to compare high and zero virus at zero time. Three new sets of variables were generated from the initial 42 variables selected using the t-test comparison between zero and 24 hours. For each set of variables the discriminant analysis program was used to develop a classification function for five groups consisting of the following:

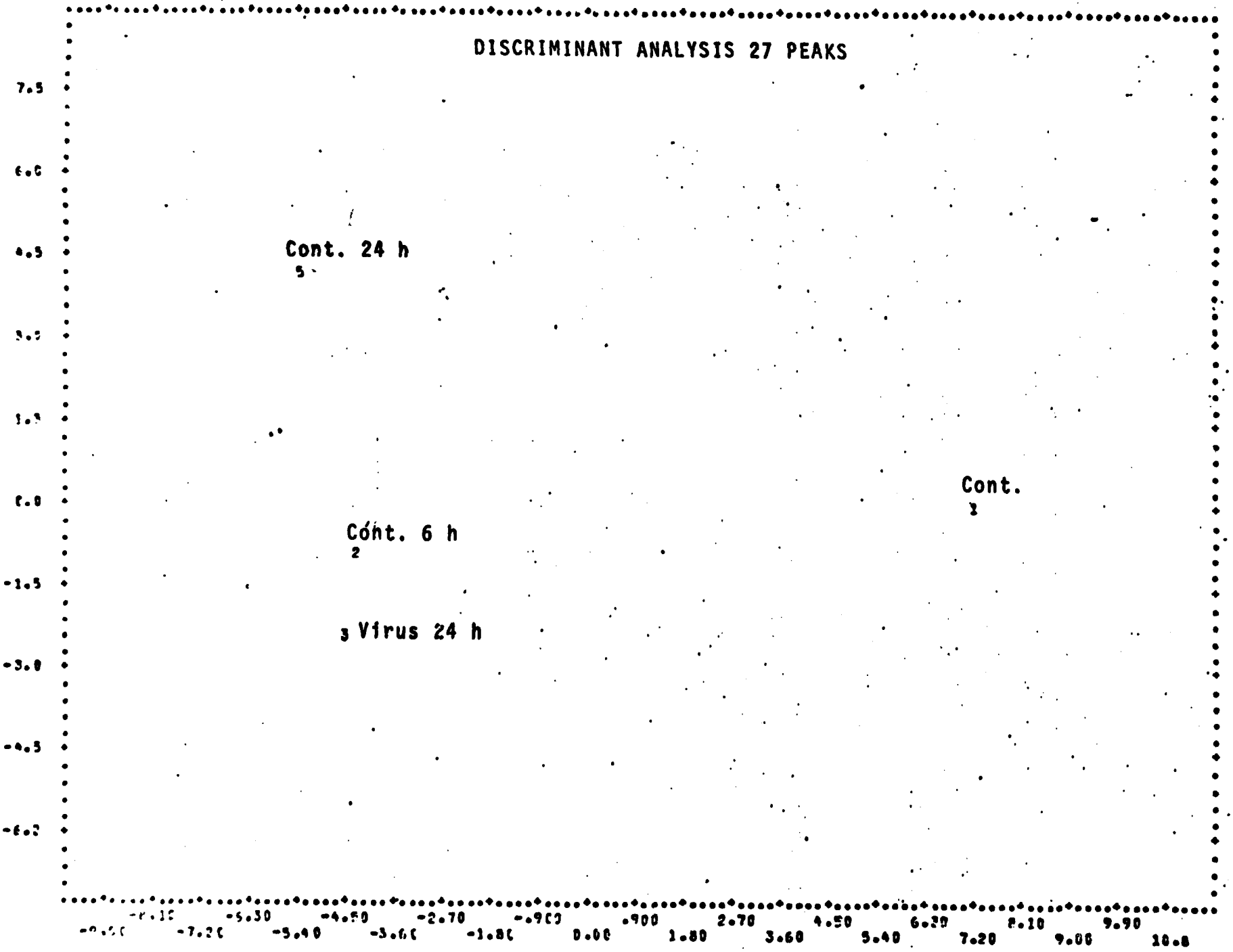
group 1: all zero hour samples

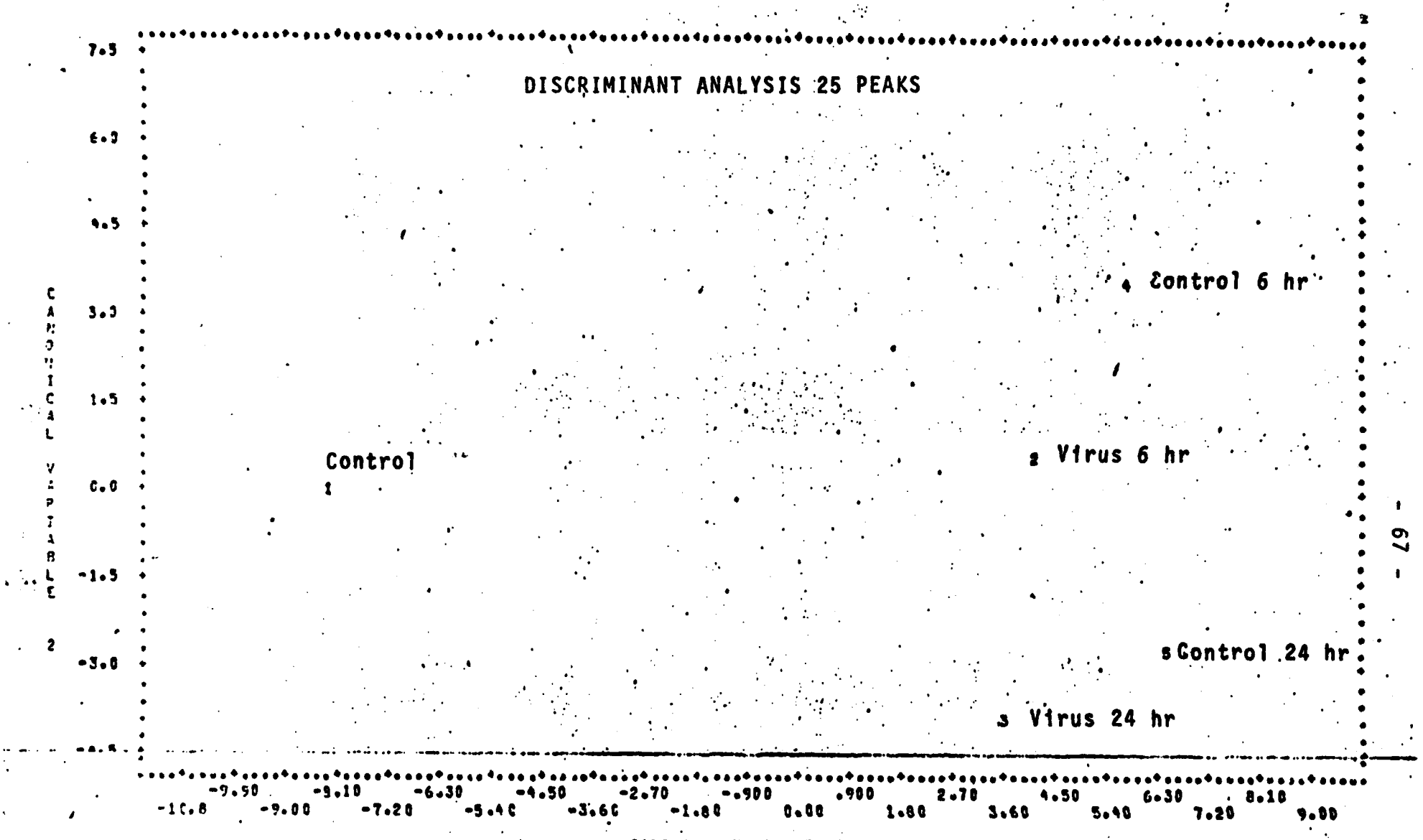
group 2: high virus, 6 hrs

group 3: high virus, 24 hrs

group 4: zero virus, 6 hrs

group 5: zero virus, 24 hours.

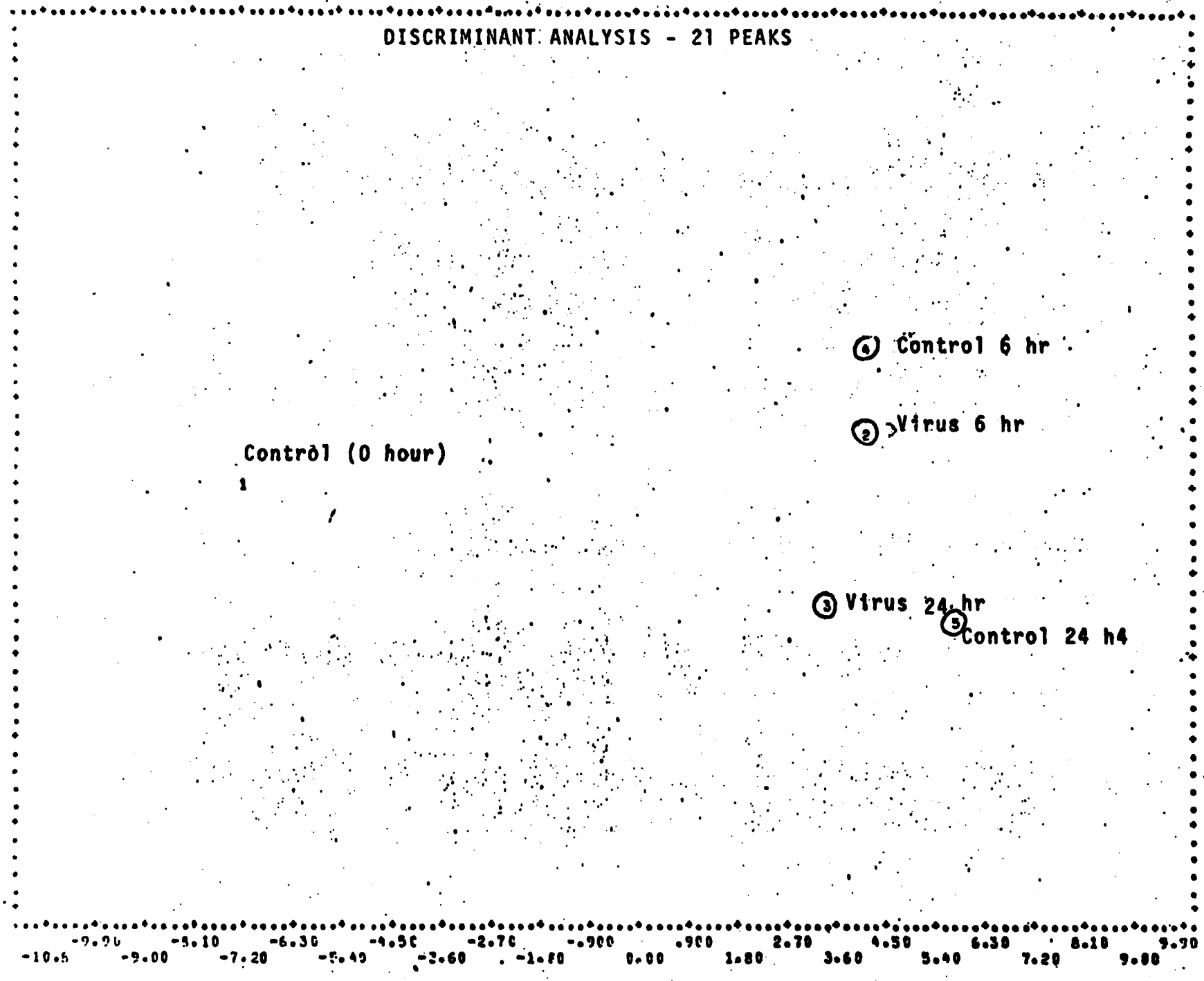




BMD07M - STEPWISE DISCRIMINANT ANALYSIS.
HEALTH SCIENCES COMPUTING FACILITY
UNIVERSITY OF CALIFORNIA, LOS ANGELES

FIGURE 33

OVERLAP OF DIFFERENT GROUPS IS INDICATED BY *



A set of 27 masses was selected from the above 42 by removing 15 masses also appearing in the zero hour test group.

These results are summarized in the canonical variable plots of the group average locations in Figures 32 through 34. The separation obtained with 27 variables (Figure 32) was increased when two additional zero time dependent masses were removed as shown in Figure 33. Reducing these 25 variable to 21 by eliminating 4 more peaks in common with the zero time difference leads to reduced separation as shown in Figure 34. Thus it is not possible to eliminate all zero time artifacts without losing some of the time and virus specific information contained in the variables removed in the third test. This compromise, however, would not exist in future experiments designed to avoid initial differences in media composition.

These preliminary results on the "metabolic profile" of tissue culture media indicates that the virus infection can be detected in vitro 6 hours following exposure. We do not know as yet whether this is the shortest time for a significant virus induced change to be demonstratable by our technique, nor do we know whether this effect of the virus is specific to a given virus species. Further study in the coming year will clarify these points.

4. PROPOSED PROGRAM OF RESEARCH FOR 1980/81

Here is our proposed program of research for the coming year, as well as an outline of some of the follow-on tasks.

A. Longitudinal Studies on Human Subjects with Infectious Disease.

We plan to continue to analyze additional samples of urine to be obtained from AFRIID from longitudinal studies on volunteers infected with Dengue fever. These experiments together with the ongoing tests on urine from a longitudinal study on sandfly fever will give us a better base to evaluate the scope of our diagnostic method. In addition we plan to collect urine samples from patients in Buffalo hospitals suffering from a number of well diagnosed infectious disease, primary viral infections. Since we found it rather difficult to obtain well documented samples from the same patient during the course of a disease, we plan to recruit medical students as project members to overcome this difficulty. We also plan to test and possibly implement a simpler and more reproducible sample preparation technique for urine samples. This technique may cut down by 30% the overall man-hour time requirement per analysis.

B. The Study of Tissue Culture Media as a Method for Early Detection of Viruses

The preliminary positive findings described in the "Technical Report" (Part III) encourage us to devote to this aspect of the program a major effort. We plan to continue the experiments on polio virus in collaboration with Dr. Howard Faden of the Buffalo Children's Hospital. We plan to extend this study to different strains of polio virus and then to other viruses. In collaboration with Dr. Nathan Woodruff of AFRIID we plan to analyze an extensive matrix of culture media comprising different viruses and different cell lines sampled at different times during the first day following infection. The simple sample sterilization and preparation techniques

developed this year will be used, and attempts will be made to modify the mass spectrometric analysis procedure to double the throughput and perform 2 analyses per hour. We hope also to extend the same procedure and identify microorganisms through their biochemical effects on bacterial culture media. Since the rapid identification of viruses is more unique we plan to give it priority over bacterial studies.

C. Continued Evaluation of Statistical Classification Procedures.

In the "Technical Report" we show that we made substantial progress in the use of different classification techniques. In the coming year we plan to devote to this aspect a full-time graduate student. At this stage of the program when the sample preparation and the mass spectrometric analysis give satisfactory results, we depend on the computerized statistical analysis to extract as much information as practical from the host of mass spectrometric data. Each of a number of potential classification techniques will be tested by a variety of challenges - such as deliberate misassignment of samples (wrong primary diagnosis) providing the program with deliberate a priori false positive and false negatives. We will then try to combine classification techniques to a sequence most appropriate for our diagnostic purpose. For instance, we may modify the P7M stepwise discriminant analysis program to become a virtual case assignment program with a fixed classification library. We may also modify the clustering program to obtain from it the centroids of individual clusters and the use of F type test of intergroup vs. intragroup variance to establish the diagnostic separation of clusters. If the results of the clustering analysis could be visually presented in a quantitative manner, it would become a preferred approach, since it does not imply any a priori grouping of the samples analyzed. If may also be possible to use the clustering analysis as a preliminary group classification and assignment procedure, followed by the discriminant analysis to optimize the

classification. The latter procedure seems in any case more effective in selection of the most significant separating variables, and therefore it would be used to select diagnostic mass peaks for structural identification. Another approach may involve the use of the discriminant analysis for the selection of a diagnostic set of peaks, which will then be used to classify new samples by the spectrum classification program existing in our INCOS software package, as described in last year's report. We also plan to test the correlation between the non-parametric Wilcoxon test and the parametric t-test on a number of systems before making a final decision on the suitability of either procedure for preselection of peaks for classification of cases.

D. Tentative Program for Follow-On Years.

The next phases of research will include the following tasks in addition to those described above:

1. Identification of metabolites found to be characteristically associated with specific infections, by the use of collisional-induced fragmentation.
2. Exploration by FI-CID of characteristic patterns in certain classes of metabolites, e.g., carboxylic acids, primary alcohols, primary amines identified by a characteristic fragment.
3. Application of multicomponent analysis to the characterization of bacteria - a topic that has not been developed further since the very first year of this program of research.
4. Development of more subtle and faster statistical analysis techniques that might be implemented on line in (practically) "real time" on our dedicated computer. This phase of the research program is essential in achieving the goal of rapid diagnosis of infectious diseases.

5. Exploration of the possibility of using other biological fluids, primarily plasma or saliva for rapid diagnosis of disease. The preparation of plasma samples for mass spectrometric analysis, which has been worked out during last year, is a first step in this direction. The examination of urine of patients with urinary infections and of pus will be undertaken to characterize these fluids through bacterial metabolites.